ENGINEERING - PSYCHOLOGY RESEARCH LABORATORY

University of Illinois at Urbana - Champaign

Technical Report EPL-78-2/AFOSR-78-5

December 1978

DEVELOPMENT AND UTILIZATION OF INTERNAL MODELS IN DYNAMIC SYSTEMS A COMPARISON OF MONITORS AND OPERATORS AS FAILURE DETECTORS

Colin Kessel

Christopher D. Wickens

Prepared for:

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE

Grant No. AFOSR 77-3380

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: AFOSR-TR- 79-0683	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle): Development and Utilization of Internal Models in Dynamic Systems: A comparison of monitors and operators as failure detectors.	5. TYPE OF REPORT & PERIOD COVERED: Interim
	6. PERFORMING ORG. REPORT NUMBER: EPL-78-2/AFOSR-78-5
7. AUTHOR(s): Colin Kessel, Christopher D. Wickens	8. CONTRACT OR GRANT NUMBER(s): AFOSR 77-3380
9. PERFORMING ORGANIZATION NAME AND ADDRESS: University of Illinois, Department of Psychology, Champaign, IL 61820	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS: 61102F/2313/A4
11. CONTROLLING OFFICE NAME AND ADDRESS: Air Force Office of Scientific Research (NL), Building 410 - Bolling AFB, Washington D.C. 20332	12. REPORT DATE: December, 1978
	13. NUMBER OF PAGES: 94
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)	15. SECURITY CLASS. (of this report): Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Autopilot
Failure detection
Internal model
Monitoring
Tracking

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The development of the internal model as it pertains to the detection of step changes in the order of control dynamics is investigated for two modes of participation: when the subjects are actively controlling those dynamics and when they are monitoring the same dynamics under autopilot control. The experiment used a transfer of training design to evaluate the relative contribution of proprioception and visual information to the overall accuracy of the internal model. The subjects either tracked or monitored the system (over)

next page

DD FORM 1473, 1 JAN 73 — EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

20.

dynamics of a 2-dimensional pursuit display under single task conditions and concurrently with a "sub-critical" tracking task at two difficulty levels. Detection performance was faster and more accurate in the manual as opposed to the autopilot mode. The development of an internal model in the manual mode transferred positively to the automatic mode producing enhanced detection performance. The cue utilization strategies of the subjects were analyzed by ensemble averaging technique and it was found that monitors of automatic systems who have had prior manual experience rely upon different perceptual cues in making their detection responses than those who have not. The proprioception channel was found to have an attention focusing role and once incorporated into the internal model this attention focusing mechanism can be used to advantage even when there is no proprioceptive feedback. A second experiment investigated in detail the effect of degrading proprioceptive information on manual mode detection. The affects of this degradation on detection performance were found to be minimal.

ENGINEERING-PSYCHOLOGY RESEARCH LABORATORY

University of Illinois at Urbana-Champaign
Champaign, Illinois

18 AFOSR 19 TR-79-0683

9 Interim rept.,

Technical Report
14 EPL-78-2 /AFOSR-78-5
11 December 1978

6

DEVELOPMENT AND UTILIZATION OF INTERNAL MODELS IN DYNAMIC SYSTEMS. A COMPARISON OF MONITORS AND OPERATORS AS FAILURE DETECTORS.

16 2313 17 A4 12 101p.

10 Colin Kessel
Christopher D. Wickens

Prepared for

AIR FORCE OFFICE OF SCIENTIFIC RESEARCH
AIR FORCE SYSTEMS COMMAND
UNITED STATES AIR FORCE

Grant No. AFOSR-77-3380
15

410 871

ABSTRACT

The development of the internal model as it pertains to the detection of step changes in the order of control dynamics is investigated for two modes of participation: when the subjects are actively controlling those dynamics and when they are monitoring the same dynamics under autopilot control. The experiment used a transfer of training design to evaluate the relative contribution of proprioception and visual information to the overall accuracy of the internal model. The subjects either tracked or monitored the system dynamics of a 2-dimensional pursuit display under single task conditions and concurrently with a "sub-critical" tracking task at two difficulty levels. Detection performance was faster and more accurate in the manual as opposed to the autopilot mode. The development of an internal model in the manual mode transferred positively to the automatic mode producing enhanced detection performance. The cue utilization strategies of the subjects were analyzed by ensemble averaging technique and it was found that monitors of automatic systems who have had prior manual experience rely upon different perceptual cues in making their detection responses than those who have not. The proprioception channel was found to have an attention focusing role and once incorporated into the internal model this attention focusing mechanism can be used to advantage even when there is no proprioceptive feedback. A second experiment investigated in detail the effect of degrading proprioceptive information on manual mode detection. The affects of this degradation on detection performance were found to be minimal.

Table of Contents

INTRODUCTION

Theoretical Overview

Advances in computer technology and their incorporation into applied settings has led to the inevitable redefinition of operator roles. Operators are slowly being moved out of the manual operating loop to become monitors of automatically controlled systems (Sheridan, 1976). Being out of the loop has a major advantage in that it has released the operator from many routine activities and in some systems has even reduced the operator's overall decision making load (Freedy et al. 1976; Rouse, 1975).

The operator is expected to stay current with the system dynamics so as to deal with unusual developments and be in a position to take over from the automatic controller should some malfunction occur. Being a failure detector while being removed from the loop has placed an added burden on the operator and changes his task sufficiently to raise a whole series of questions about the nature of the changed role and its impact on overall system performance. This study has examined a number of basic variables that have both theoretical and practical implications for the changed operator roles.

In an automatic task, the system response is computer controlled and the human operator (the monitor) sees the system output as presented by situation displays and the overall system performance. In this mode the typical monitor has only visual cues upon which to base decisions about system performance. In the manual mode, however, the human operator is now a controller in the loop and interacts with the system via some proprioceptive motor control. The controller therefore, has an added information channel, his knowledge of commands delivered to the

system revealed from proprioceptive cues, which can be used to aid in decision making about the state of the system. It is via this channel, together with the visual, that the controller can test out strategies and develop an adaptation which "encompasses the complex matter of optimization of the manual control loop on the basis of various control objectives" (Young, 1969, p. 293).

When performing a dynamic continuous task like pursuit tracking, it is assumed that operators develop an internal representation of the dynamics of the system (Veldhuyzen and Stassen, 1976; Pew, 1974). This is true for both the manual mode in which the controller is in the loop as well as the automatic mode in which the monitor is out of the loop monitoring an automatic controller.

The internal model refers to the internal representation of the state of the normally operating system as depicted by the expected values of state variables and their expected variability (see Fig. 1). Veldhuyzen and Stassen (1976) define the internal model as simply the "internal representation of the knowledge the human operator has" (p. 158). The internal model of the system serves as a basic premise for a number of manual control models such as the quasi-linear model (McRuer and Krendel, 1974) and the optimal control model (Kleinman et al., 1971). While the quasi-linear model does not explicitly specify the existence of an internal representation of the system it does assume that the operator internalizes system dynamics and uses this knowledge in order to produce some stability in controlling the system. The optimal control model, however, has built specific mechanisms into the model to account for the internalization of system dynamics.

After identifying a number of limitations in the use of the internal model concept as it has been used in control theory Veldhuyzen

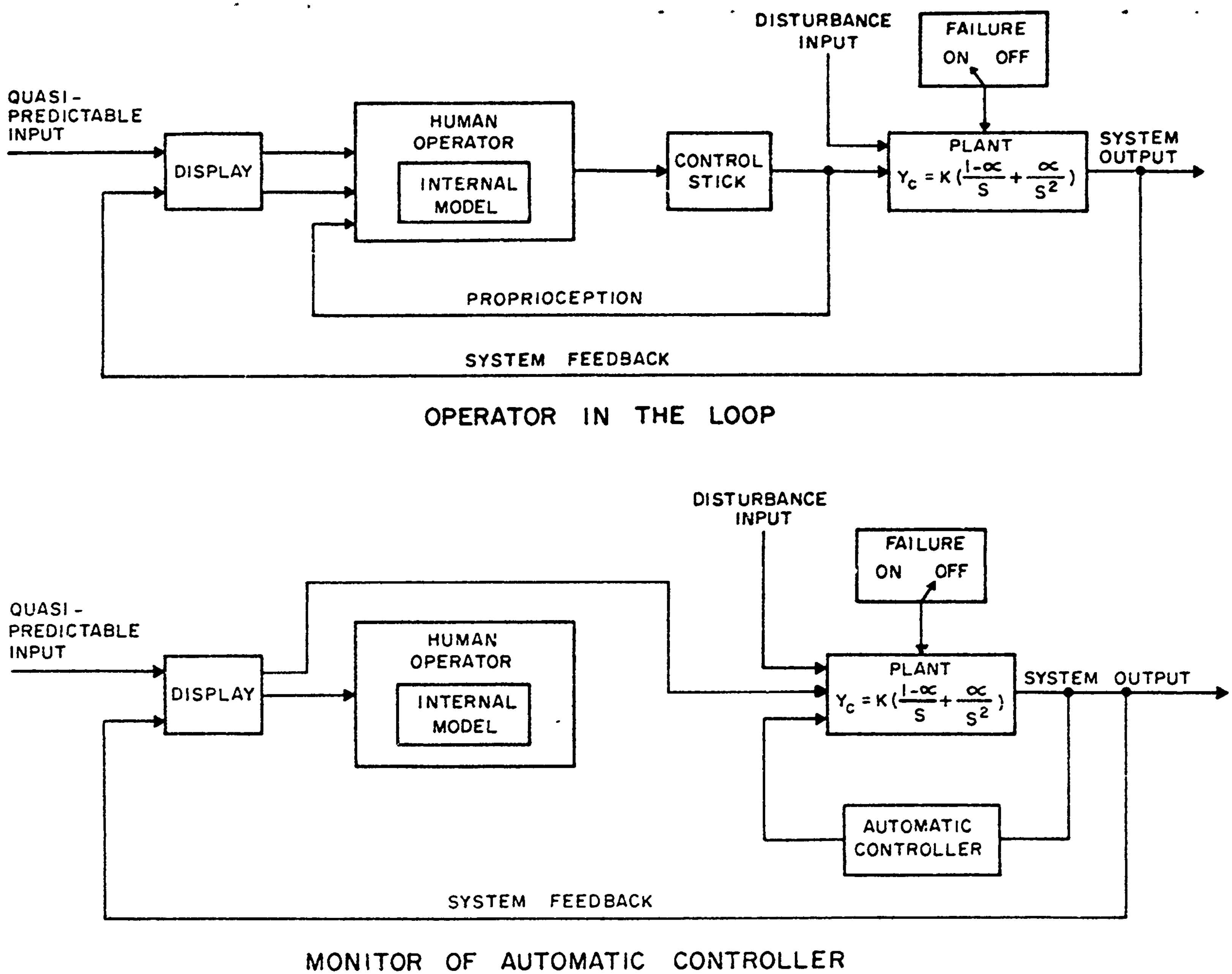


Figure 1. Schematic representation of failure detection process in control mode(top),monitor mode(bottom).

and Stassen (1976) conclude that "the study of the meaning of the Internal Model concept is of great importance in understanding human performance because the monitoring, decision making, predicting or extrapolating and planning activities of human beings are all based on an Internal Model" (p. 159).

The internal model can be viewed in terms of the way it affects overall performance in a particular interactive situation or it can be seen in relation to the way it is developed. The development of the internal model is a continuing ongoing process and continues during all the stages of interaction between the operator and the system. The performance of the operator, on the other hand, can only be measured in terms of the system's overall output and efficiency.

It is this dichotomy between learning (the development of the model) and performance (the sensitivity of the model to system changes) that is the subject of the current research project. Three fundamental research issues have been identified: 1) to examine the role of a separate set of information channels; 2) to determine the relative sensitivity to system changes of different types of internal models; and 3) to establish whether the way the internal model is developed influences its subsequent sensitivity to system changes.

While the first issue has been studied before (Wickens & Kessel, 1977; 1979a) the second and third represent an original approach to this whole problem. By employing a between subject design and utilising a transfer of training technique (see p. 9) this study was able to examine how different training schedules based on the separate development of independent internal models, each based on a different set of information channels contributed to the relative sensitivity of these internal models to system changes. The development of the internal

model therefore must be seen as a function of the number of information channels available, the nature of the information channels and their relative independence of learning from other internal models. Particular attention was paid to the relative contribution of visual and proprioceptive feedback and a theoretical model was developed to account for the contribution of each to the formulation of the internal model of the system.

While the internal model is recognized as being important there is a basic problem in utilizing this concept and this relates to its relative inaccessibility to experimental manipulation. One way of gauging the current status of the internal model is through inferences from the relative sensitivity to system changes of controllers as opposed to monitors. This methodology therefore forms the basis for examining all the theoretical issues relating to the internal model described above.

The detection of a failure or change in the characteristics of a dynamic system requires that the detector has available three basic elements: (1) an internal representation of the state of the normally operating system--the expected value of state variables, and their expected variability (Veldhuyzen and Stassen, 1976; Pew, 1974; Rouse, 1977); (2) a channel, or set of channels, of information concerning the <u>current</u> state of the system. Failures are detected when the information concerning the current system state is assessed to be sufficiently deviant from the representation of normal operation to warrant a decision. The decision process involved has been assumed to involve the application of some statistical decision rule (Curry and Gai, 1976); (3) the options of testing hypotheses about the nature of the dynamics by introducing signals into the system (Hess, 1978). For a detailed

theoretical analysis of these three elements see Wickens and Kessel (1979a).

Problem Definition

Both the experimental evidence (Young, 1969; Wickens & Kessel, 1977; 1979a) and the theoretical analysis above provide considerable support for the conclusion that the manual mode is superior to the automatic mode in failure detection. It is not clear however, whether this manual superiority resides at the level of performance (i.e., the nature and number of cues available during the failure detection task) or because the internal model developed in the manual mode is more sensitive and therefore better at failure detection.

The above theoretical analysis identified two attributes that seem to facilitate failure detection in the manual mode. The inclusion of the proprioceptive channel of information not available in the automatic mode and the option of developing a strategy of hypothesis testing. It was furthermore argued that the existence of these two factors would not only enhance performance but would combine in the learning phase to help establish a more stable and more sensitive internal model.

In comparison the automatic mode was characterized by two attributes that could facilitate detection: a greater "strength" of the visual signal (since adaptation by an automatic controller does not take place) and a lower level of workload. These two attributes however do not necessarily provide the monitor with much information to help develop a less variant internal model during the learning phase.

A characteristic of all the previous research in this area (Young, 1969; Ephrath & Curry, 1977; Wickens & Kessel, 1977a; 1977b) is that the learning factor could not be separated from the performance factor and no conclusion could be reached about the importance of the learning

stage in determining the sensitivity of the internal model to system changes. This limitation is based mainly on the utilization of a within subject design that by definition does not allow for a separation of learning from performance.

This study has employed a between subject transfer of training design that will enable the comparison of the two modes of participation each based on a separate and unique internal model. This design enables a greater understanding of the relative importance of the number and nature of cues available, and the importance of adaptation and the significance of hypothesis testing strategies in developing internal models. Should all these factors be important it is projected that the manual mode will prove to develop a more sensitive and less variable internal model than the automatic mode.

This design will also establish how these internal models interact with one another. By comparing these results, based on the between subject design with results of the previous experiments based on a within subject design the relative independence of the internal models as measured by the amount of interference can be guaged.

Once the separate internal models have been developed for each mode of participation their relative transfer can be measured by determining how an internal model developed in one mode can be utilized in the development and subsequent performance in the other mode. It is hypothesized that internal models can be characterized in terms of how they were developed and how they are used. An internal model based on a "richer" set of cues (as in the manual mode) will prove to be more stable and more sensitive to system failures than a model based on fewer cues. The utilization of an internal model is a function of both how the model was developed and the number of cues available at the time of

utilization. In the manual to automatic transfer situation therefore it is expected that the automatic group will be able to utilize to advantage an internal model developed during the prior "richer" manual mode.

Finally, it is argued that any advantage of monitoring over controlling attributable to workload differences might itself be dissipated as the competition for attentional resources is increased by imposing concurrent tasks. This interplay of factors, and their manipulation facilitates a clearer identification of the nature of the failure detection task and allows predictions to be formulated concerning the differential effects of variables such as workload or control adaptation on detection performance.

This study therefore addresses the basic question of the nature of the internal model, examines the relative contribution of information channels in its development, and measures the relative sensitivity of different internal models to system changes, utilizing failure detection performance as the operational definition of internal model strength.

EXPERIMENTAL OVERVIEW AND PREDICTIONS

This research project was designed to answer three main questions. The first relates to the differences in development of the internal models of operators and monitors and their subsequent utilization in a failure detection task. The second question is directed to the impact of concurrent task workload while the third question relates to the role of proprioceptive cues in failure detection. Each question was studied by a separate experiment. The overall experimental design is schematically represented in Figure 2.

Experiment 1

The first question defined above was studied by using a transfer of training technique. The basic transfer of training design is presented in Table 1. This design employs two experimental groups and two control groups. By holding all the experimental conditions equal except the mode of participation it is possible to compare the relative contribution of visual as opposed to proprioceptive cues to the failure detection performance.

In method one a comparison is made between the experimental group which transfers an internal model formed during training on the manual mode to one based on the automatic mode with a control group that experiences the automatic mode only. This therefore enables a comparison of the relative contribution of prior learning using both visual and proprioceptive cues on subsequent failure detection with prior learning when only visual cues were available. The second method reverses this condition and examines a transfer of information based on visual cues only to one based on both visual and proprioceptive cues.

Should the above theoretical analysis of the relative contribution

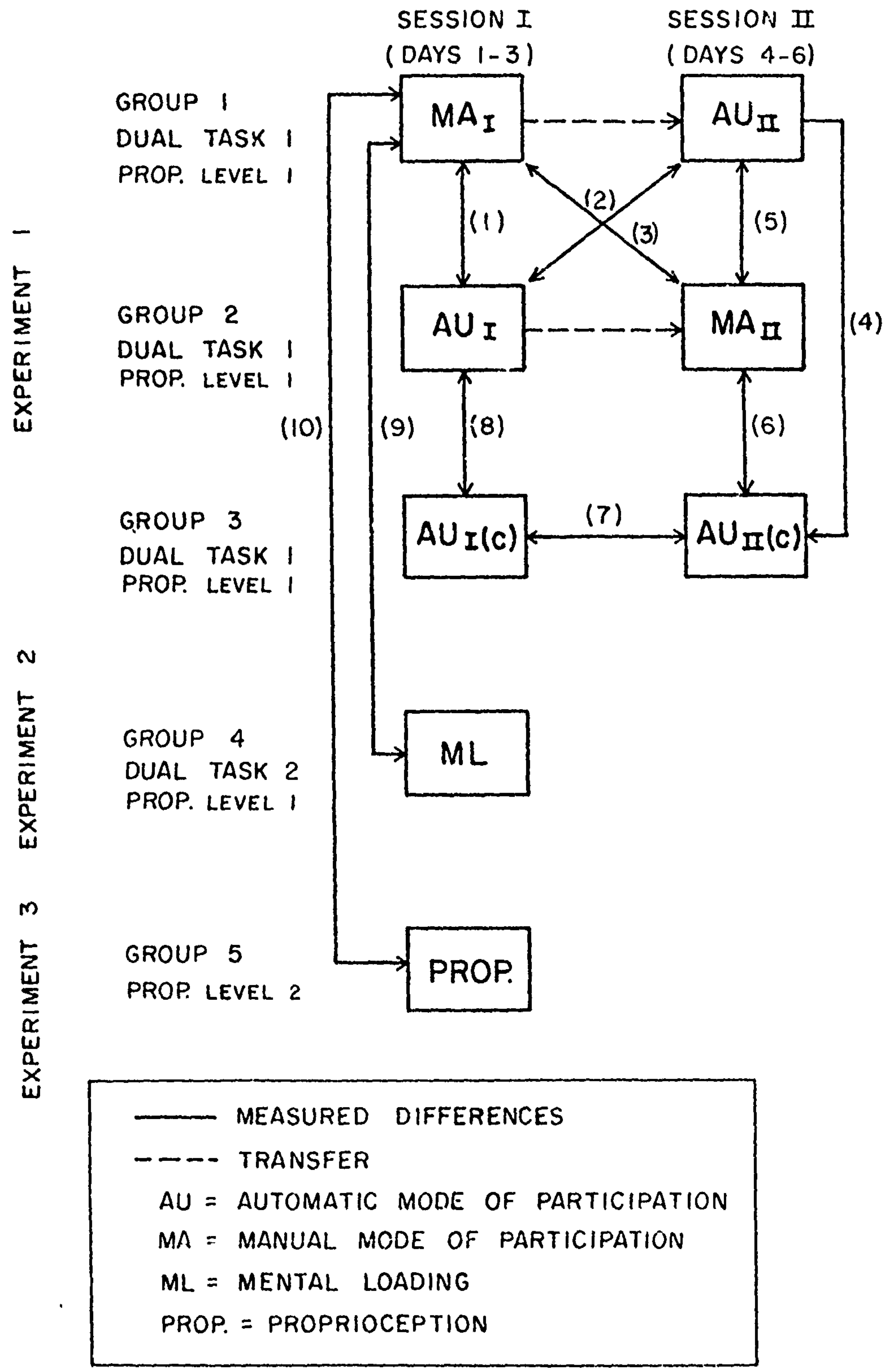


Figure 2. Experimental design and expected relationships between experimental conditions.

		Order of Presentation	
		First Series	Second Series
Method-1	Experimental Group	Task A (MA)	Task B (AU)
	Control Group	Task B (AU)	Task B (AU)
Method-2	Experimental Group	Task B (AU)	Task A (MA)
	Control Group	rest	Task A (MA)

AU = Automatic Mode

MA = Manual Mode

Table 3

Transfer of Training Design

of visual and proprioceptive cues in internal model development hold true, the transfer in method one should enhance subsequent failure detection therefore demonstrating that an internal model built up on both proprioceptive and visual cues will facilitate failure detection in a task based solely on visual cues. It also follows from the above argument that there should be relatively minimal transfer from the automatic mode, which is based on visual cues only to the manual mode which utilizes both visual and proprioceptive cues. It is expected therefore that an internal model built up on both visual and proprioceptive cues will be more sensitive and less variable than one built up on visual cues alone. This theoretical analysis has therefore led to a number of specific predictions.

With reference to the specific relationships depicted in Figure 2 and referred to by the numbered arrows the following predictions can be made:

(1) Expect a significant difference in detection performance between the manual control group (MA_I) and the automatic group (AU_I). This comparison (number 1 in Figure 2) represents a replication of the Wickens and Kessel (1977) study, however the between subject design of this study is expected to produce greater differences than in the Wickens and Kessel study. It is expected that these manual-automatic differences will be repeated for the post transfer groups (i.e., comparison no. 5 for groups MA_{II} - AU_{II} and comparison no. 6 for groups MA_{II} - $AU_{II(c)}$).

(2) Expect a significant difference in detection performance between AU_I and AU_{II} (comparison no. 2) and AU_{II} - $AU_{II(c)}$ (comparison no. 4) which will result from the positive transfer effect from the MA_I group to AU_{II} group. This positive transfer can be attributed to the AU_{II}

groups ability to utilize to advantage the internal model, based on both visual and proprioceptive cues developed in the MA_I condition.

(3) Expect no difference in tracking or detection performance between MA_I and MA_{II} (comparison no. 3). The MA_{II} will not be able to benefit from an internal model based on visual cues and developed in the AU_I condition and will therefore be forced to develop a new and relatively independent internal model.

(4) No significant differences are expected for comparisons number 8 (AU_I - $AU_{I(c)}$), or number 7 ($AU_{I(c)}$ - $AU_{II(c)}$). These comparisons are essentially between control groups and both between AU conditions therefore no advantage is expected for any one group.

Experiment 2

The second question relates to the role of concurrent task workload and was studied by comparing the impact of two different loading tasks on the primary tracking and failure detection task. This experiment is discussed in greater detail in Wickens and Kessel (1979b).

Experiment 3

The third question defined above relates to the role of proprioceptive cues in the development of the internal model. This question was studied by comparing a group that tracks with an isotonic control stick, (PROP.), in which all spring resistence was removed, with the control group (MA_I) in experiment 1 that operated with the normal spring loaded control stick (for description see p. 20).

The theoretical discussion in the Introduction has led to a specific expectation: 1) due to the importance of proprioception it is expected that the group with the stronger proprioceptive cues (MA_I) will produce a higher level of detection than the group with the degraded

proprioceptive cues (PROP.) (comparison no. 10 in Figure 2).

HYPOTHESES

From the above theoretical discussion four main research hypotheses can be formulated:

(1) The manual mode of participation will produce better detection of failures in a dynamic system than the automatic, monitoring mode of participation.

(2) The manual mode superiority can be accounted for by the existence of a more sensitive internal model developed by both proprioceptive and visual cues as opposed to the weaker internal model in the automatic mode that is based on visual cues only. A transfer of training paradigm can be used to demonstrate the relative superiority of the manual mode over the automatic mode by showing how the monitor in the automatic mode utilizes to added advantage an internal model that was previously developed in the manual mode.

(3) Workload will interact with failure detection performance but the interaction will be dependent upon the nature of the concurrent task. A central processing loading task will interfere more severely with the failure detection task while the manual side task will have greater impact on overall tracking performance. (This hypothesis is dealt with in Experiment 2, as reported in Wickens and Kessel, 1979b.)

(4) The nature and quality of proprioceptive feedback will have a direct impact on the overall sensitivity of the internal model. The stronger the proprioceptive feedback the more sensitive the internal model to system failures.

EXPERIMENTAL INVESTIGATION

The following section presents a definition of the basic elements of the experimental paradigm. The overall paradigm was designed to examine the operator's failure detection performance as a joint function of the participatory mode, the means of development of the internal representation and the workload demands imposed by side tasks of varying difficulty and varying demand. Three experiments were run. In converging upon the particular experimental configuration that was used, a number of decisions were made. It was decided, for example, to employ pursuit rather than compensatory display dynamics to enable a clearer separation, in the automatic mode, of inputs due to target following from those due to disturbance functions.

The selection of failures to be used--step changes in system order--was dictated by a desire both to simulate plausible events in a real world environment (loss of stability augmentation) and also to produce failure "signals" that would not be so obvious that their detection would be guaranteed. Finally, the relatively high frequency of failure occurrence, five per 2 1/2 minute trial on the average, was selected for the purpose of generating enough data to make reliable estimates of performance while acknowledging that this frequency departs from the much lower expectancy of failures in operational settings (Earing, 1977).

It should be noted that these experiments represent a continuation of five preliminary investigations whose joint functions were to shape the formulation of the paradigm employed here, select the appropriate level of experimental variables, and perfect analysis and measurement techniques.

Apparatus

The basic experimental equipment included a 7.5 x 10 cm. Hewlett Packard Model 1300 CRT display, a spring-centered, dual-axis tracking hand control (with an index-finger trigger) operated with the dominant hand and a spring-centered finger control operated with the other hand. The two hand controls were maintained at a constant level of resistence. A Raytheon 704 16-bit digital computer with 24 K memory and A/D, D/A interfacing was used both to generate inputs to the tracking display and to process responses of the subjects. The subject was seated on a chair with two arm rests, one for the tracking hand controller and one for the side-task finger controller. The subject's eyes were approximately 112 centimeters from the CRT display. The overall display subtended 1.5 degrees--therefore falling within foveal vision.

Pursuit Tracking Task

The primary pursuit-tracking task required the subject to match the position of a cursor with that of a target which followed a semi-predictable two-dimensional path across the display. The target's path was determined by the summation of two non-harmonically related sinusoids (.05 and .08 Hz) along each axis with a phase offset between the axes. This produced a target that moved along a path producing a randomly appearing figure eight. The position of the following cursor was controlled jointly by the subject's control response, produced by manipulating a hand controller with the dominant hand and by a band-limited forcing function with a cutoff frequency of .32 Hz for both axes (see Figure 3). Thus the two inputs to the system were well differentiated in terms of predictability, bandwidth, and locus of effect (target vs cursor). The control dynamics of the tracking task were of the form $Y = K \left(\frac{1-\alpha}{S} + \frac{\alpha}{S^2}\right)$ for each axis, where is the

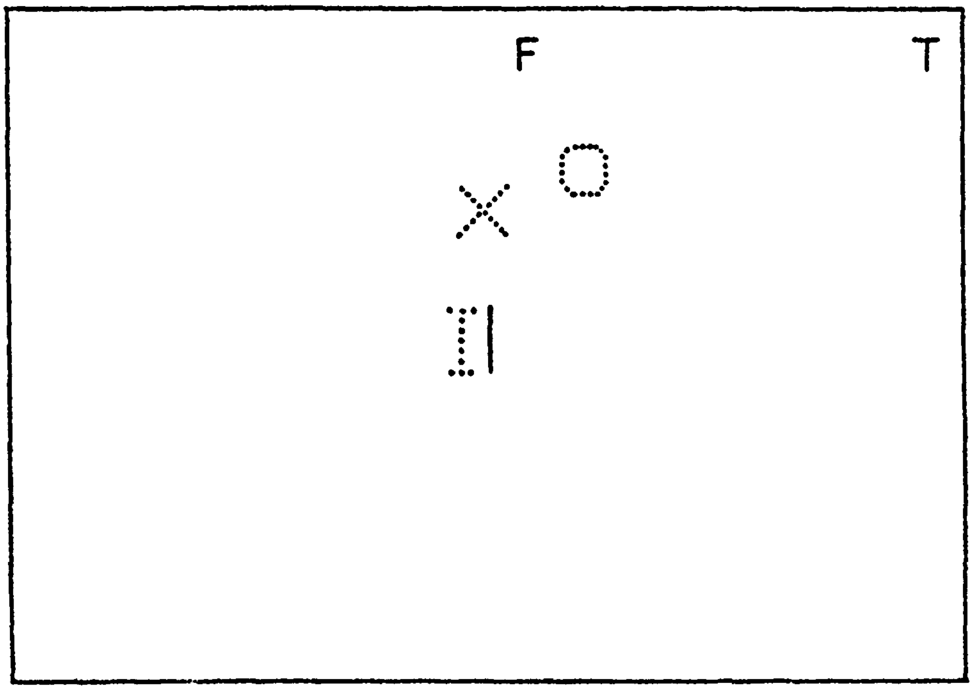


Figure 3 Subject's display: The X and O represent, respectively, the tracking cursor and target for the two-dimensional pursuit tracking task; the I in the center is the error cursor for the compensatory Critical Task; F and T indicate, respectively, the occurrence of a failure during training sessions and the onset of the trigger.

variable parameter used to introduce changes in the system dynamics. K is the gain of the stick, while S refers to the LaPlace operator. The division of S corresponds to an integration of the input over time. These changes, or simulated fialures, were introduced by step changes in the acceleration constant α from a normal value of .3, a mixed velocity and acceleration system with a high weighting on the velocity component, to α = .9, a system that approximates pure second-order dynamics, and required the operator to generate considerable lead in order to maintain stable performance.

Side Tasks

(a) Critical Task (Experiment 1)

The first side task termed a Critical Task by Jex, McDonnel and Phatac (1966), was displayed horizontally at the center of the screen and required the subject to apply force to the spring-loaded finger control in a left-right direction to keep an unstable error cursor centered on the display. The control dynamics of this task were of the form $Y = \frac{K\lambda}{S-\lambda}$. These dynamics formed an unstable positive feedback loop that drove the error cursor to the edge of the display at a velocity proportional to the error and to the parameter λ. The difficulty of the Critical Task was therefore controlled by adjustment of λ. Assuming constant performance is maintained on the critical task, manipulation of λ served to vary the extent of processing resources demanded by the task. Two values ($\lambda = 0.5$ and $\lambda = 1.0$) were employed on different dual task trials.

(b) Memory Transformation Task (Experiment 2)

Experiment 2 employed a different side task, one involving a memory transformation task, and is discussed in Wickens and Kessel (1979b).

Levels of Proprioception (experiment 3)

The strength of the proprioceptive feedback was manipulated by changing the resistence of the dual axis hand control. Two levels of resistence were used. In the normal operating state the resistence of the springs was set at 520 grams at maximum displacement. In the isotonic stick mode of operation the spring resistence was at zero. This was achieved by disconnecting the springs from the normal stick. All other parameters of the tracking task remained the same therefore requiring the subject to make the identical movements of the control stick to achieve equivalent levels of tracking.

Equating the Experimental Groups

In order to ensure that the subjects in all the experimental groups started the experiment with equivalent tracking proficiency each subject was randomly assigned to one of the five groups and using their non-dominant hand performed 10 one minute trials in the Critical Task using high levels of lamda (= 2.5 and 3.0). These high levels of lamda ensured that the Critical Task was at a very unstable level demanding a high degree of skill. The average R.M.S. error on these ten trials were used to ensure that the initial equality in tracking performance of the five groups was achieved.

EXPERIMENT 1

Subjects:

The subjects were 18 male university students. Subjects were payed a base rate of $2.50 per hour but could increase their overall pay by maintaining a high level of detection performance (see payoff schedule below).

Payoff Schedule:

In order to ensure that all the subjects maintained a fairly constant level of incentive throughout the whole experiment subjects were payed a base pay plus 50 cents for maintaining a constant level on the side task plus 2 cents for every correct detection and they lost 1/2 cent from their total for every false alarm. Using this payoff schedule subjects could maximise their pay to $3.50 per hour, per experimental session.

Experimental Design:

Experiment 1 employed a transfer of training design (see Tables 1 and 2) to examine the relative contribution of information channels to the development of the internal model. Table 1 represents the overall theoretical design, requiring four groups--two transfer groups and two control groups, while Table 2 represents the actual design used. From this table it can be seen that only 3 experimental groups were needed since the first three sessions of group 1 served as the manual control group for experimental group 2. There were six subjects in each group and each subject participated in six consecutive experimental sessions. Each session lasted 1 1/2 hours and took place on consecutive days. Subjects in group 1, for example, participated in 3 manual (MA) sessions and then 3 automatic (AU) sessions. The first day of each condition was

	Session I			Session II			
	Day 1 Training	Day 2 Experimental Sessions	Day 3	Day 4 Training	Day 5 Experimental Sessions	Day 6	No. of Subjects
Group 1 * Manual to Automatic	MA_I	MA_I	MA_I	AU_{II}	AU_{II}	AU_{II}	6
Group 2 Automatic to Manual	AU_I	AU_I	AU_I	MA_{II}	MA_{II}	MA_{II}	6
Group 3 Control-Automatic	$AU_{I\ (c)}$	$AU_{I(c)}$	$AU_{I(c)}$	$AU_{II(c)}$	$AU_{II(c)}$	$AU_{II(c)}$	6

*NOTE: The first three sessions of group 1 served as the manual control group for group 2 in experiment 1 and also served as the control group in experiments 2 and 3.

Table 2

Experiment 1 - Transfer of Training

a training session (see Table 3) and the next two days were experimental sessions (see Table 4).

Experimental Procedure:

Each subject participated in six sessions (see Table 2). On days 1 and 4 in a training session and on days 2, 3, 5 and 6 in experimental sessions.

a) Training Sessions:

During the training session (see Table 3) the subject participated in a number of different types of trials designed to provide experience and practice in both the normal operating mode and the failed condition. The subject performed in either of the two modes of participation, the manual tracking mode (MA) or automatic mode (AU). In the MA mode the subject performed the tracking manually, while in the AU mode his role in the control loop was replaced by automatic controller dynamics consisting of a pure gain and time delay. The open-loop gain was set at a constant value for all subjects while the time delay and the disturbance function were set at values (time delay = 540 ms[1] and disturbance cutoff frequency = .4) which were determined in a pretest experiment. This procedure ensured that the AU tracking performance was equivalent to the overall expected MA single task tracking performance. These values of time delay and disturbance frequency were maintained throughout the rest of the experiment. Each trial, MA or AU lasted 150 seconds.

After completing three training trials in the MA or AU mode only the subject performed four single task trials in the Critical Task. In the Critical Task the subject was instructed to apply force to the finger controller along the X-axis to keep the "I" balanced in the

Day 1

Order	Number of Trials	Type of Trial
1	3	AU or MA
2	2	Easy Dual Task Only
3	2	Difficult Dual Task Only
4	2	AU or MA + Easy Dual Task
5	2	AU or MA + Difficult Dual Task
6	2	AU or MA in Failed Condition
7	2	AU or MA + Easy Dual Task) + announced
8	2	AU or MA + Difficult Dual Task) failures
9	3	AU or MA experimental trials
Total	20	

Day 4

1	3	AU or MA
2	2	AU or MA + Easy Dual Task
3	2	AU or MA + Difficult Dual Task
4	2	AU or MA in Failed Condition
5	2	AU or MA + Easy Dual Task) + announced
6	2	AU or MA + Difficult Dual Task) failures
7	6	AU or MA experimental trials
Total	19	

Table 3

Experiment 1 and 2 - Training Day Design

center of the vertical line in the center of the display (see Figure 3).

The subject then performed 4 trials in which the two tasks were conducted together. After completing these training trials the subject was told that a certain number of changes would be introduced into the system and that he would be examined on his ability to detect these changes. A decision was recorded by pressing the trigger on the control stick. Pressing the trigger presented a "T" on the screen (see Figure 3) and returned the system to normal operating conditions. If the change was not detected, the system returned to normal after 6 seconds. This interval was determined by a pretest (see analysis section and Figure 4). The return to normal was achieved via a gradual ramp and took 4 seconds to complete. This precaution was made necessary by the tendency observed in pretests for subjects to view a step return to normal as the change to be detected.

To provide experience with the failed condition (i.e., the higher acceleration in the control stick), the subject received two trials in which he tracked (or viewed the automatic controller tracking) only in the failed condition. Four demonstration trials were then presented in which the subject or the computer tracked in the regular condition, but the onset of each failure was cued by the presentation of an "F" on the screen (see Figure 3). The subject was instructed to press the trigger to return the system to normal only upon the detection of the nature of the change.

After completion of these seventeen training trials, three experimental trials were then presented with both MA or AU and the dual task. The subjects were told to detect as many changes as possible as quickly as possible. The number of changes on each trial was not announced though the subject was told that no change would occur during

the first 15 seconds of each trial. The presentation of the change was generated by an algorithm that assured random intervals between presentations and allowed the subject sufficient time to establish baseline tracking performance before the onset of the next change. Task logic also insured that changes would only be introduced when system error was below a criterion value. In the absence of this latter precaution, changes would sometimes introduce obvious "jumps" in cursor position. During these three trials subjects received feedback about their overall detection performance, and their performance on the side task.

b) Experimental Sessions:

After each day of training every subject then conducted 2 experimental sessions (for the specific design of each session see Table 4). After four refresher trials in the AU or MA modes with side task and demonstrated failures the subjects conducted 15 experimental trials (5 in each experimental condition).

The subject was instructed to "do the side task as efficiently and accurately as possible." Even for trials on which this task appeared to be difficult, subjects were instructed to try to maintain a standard level of performance on the side task. After each trial the subject received feedback about his performance on the side task and was encouraged to maintain his overall level of performance. The instructions, feedback and payoff schedule therefore clearly defined the side task as the loading task while allowing the tracking and detection tasks to fluctuate in response to covert changes in available attentional resources. In this manner, workload demands were experimentally manipulated, rather than being passively assessed.

Training Trials	1) MA or AU only 2) MA or AU + Side Task 3) MA or AU + Side Task + announced failure 4) MA or AU + Side Task + announced failure	

		Work Load		
		Single	Dual	
			Easy	Hard
Experimental Trials	Mode of Participation MA or AU	1) 4* failures 2) 6 failures 3) 4 failures 4) 6 failures 5) 5 failures	1) 4 failures 2) 6 failures 3) 4 failures 4) 6 failures 5) 5 failures	1) 4 failures 2) 6 failures 3) 4 failures 4) 6 failures 5) 5 failures

*Order of trials were blocked and randomized within each block.

NOTE: There were 25 failures per experimental condition per session therefore a total of 50 failures on the two days of data collection per experimental conditions per subject.

Table 4

Exp. 1 and 2 - Experimental Design Per Session

(Days 2, 3, 5, 6)

Data Collection and Analysis

a) Detection Performance:

Following the procedure outlined by Watson and Nichols (see Appendix A), it is necessary initially to specify the interval following each failure signal to be designated as a "hit" interval. The data from a number of pretests, presented in Figure 4, indicated that the distribution of subject responses, following signal occurrence, showed a peak at around three seconds and reached a relatively stable baseline by six seconds following a failure. Therefore, 6-second intervals were defined as hit intervals, and the measure P(HIT) is simply the number of detection responses falling within the interval divided by the total number of system failures. The remaining duration of the trial (150 seconds - 6 x 4 or 6 x 5 or 6 x 6 depending on whether 4, 5, or 6 failures were presented on the trial) is similarly subdivided into 6-second false alarm intervals. The measure P(F/A) is computed as the number of false alarms divided by the number of false-alarm intervals.

Because of the relatively small number of signals presented, and the questionable applicability of the formal signal detection theory assumptions to the current data, the nonparametric measure of the area under the ROC curve, P(A), was employed as the bias-free measure of sensitivity (Green & Swets, 1966; Egan, 1975). For a sensitivity measure the area under the ROC [P(A)] was employed rather than d′ because the former measure is more robust to violation of distribution assumptions and to small numbers of signals employed here (Green & Swets, 1966). Values of this measure were computed from the P(HIT) and P(F/A) data by reference to tables in McNicol (1972). This measure produced a score varying from 0 to 1.0 for which 0.5 represents chance performance and 1.0 represents perfect accuracy. Both the P(A) measure

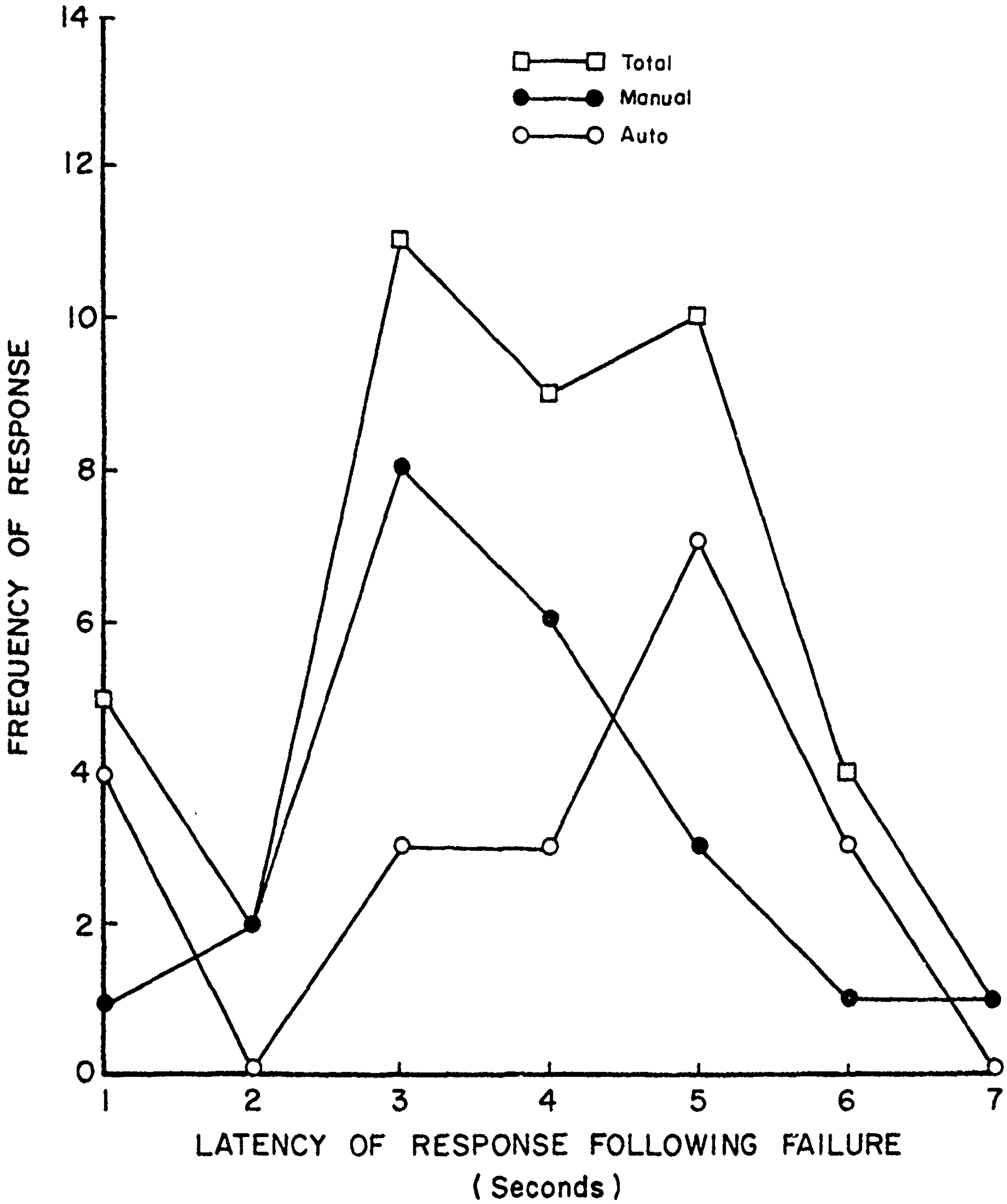


Figure 4. Latency distribution of detection responses following failure.

and the mean and standard deviation of detection latencies were computed at the end of each trial. The overall performance measures (P(HIT), P(F/A) and latencies) were pooled over all trials within each experimental condition on each day for each subject. The probability and latency scores were therefore based on 25 failures in each experimental condition. The latency score was therefore the average of the responses made while the detection score was the probability of detection out of a total of 25 possible failures each session. The P(A) measure for each subject was based on the P(HIT) and P(F/A) for 50 failures (i.e., two days of data collection). The P(A) for the group scores were based on the average of the P(HITS) and P(F/A)'s for each subject.

The P(A) measure and the latency measure were then plotted in the form of a joint speed-accuracy measure depicted in Figure A-1 (see Appendix A). "Good" performance is represented by points lying on the upper left part of the scale, in the region of a fast accurate response. Performance was quantified by projecting the joint locus obtained onto the performance axis. The performance scale is computed as [10 times P(A) - latency] and will be called the "derived performance score." This procedure produces a performance scale that ranges from zero at chance level of accuracy with a latency of 5 seconds to 10.0 for perfect detection with a zero reaction time.

The units assigned to this performance index are clearly arbitrary but are based on the finding that the overall variability (standard deviation) of the raw latency scores were found to be about 10 times the variability of the P(A) measure. Furthermore, it was observed that the clear linearity between accuracy and latency would allow this parsimonious presentation of the data without any significant changes in

the results or relationships between the experimental groups.

Since this study involves both a between and a within subject design a mixed-model 3 way analysis of variance was used to test the main hypotheses pertaining to detection accuracy and latency (see Table 5 and Results section below).

The transfer of training analyses employed a comparison both within subjects and between groups of the relative transfer from sessions 1-3 to sessions 4-6. Positive transfer is inferred if subjects who have previously learned one method (for example via the manual mode) perform significantly better in failure detection on the second method (the automatic mode) than do equivalent subjects in the appropriate control group.

Finally a detailed analysis of the distribution of response latencies was conducted. This method used in the Wickens and Kessel (1977) study involves calculating the cumulative probability distributions relative to the number or probability of failures detected as a function of latency after failure. Lappin and Disch (1972) have argued that a similar representation of his reaction time data--the latency operating characteristic (sometimes referred to as the cumulative accuracy function)--may provide evidence bearing on the time dependent processes involved in detection: the integration of perceptual evidence over time.

b) Tracking Performance:

The following analog signals were sampled every 60 msec and stored on digital tape for later data analysis: tracking vector error[2], vector stick position, and Critical Task error. In addition, on a fourth channel, the occurrences of failures and responses were recorded. At the end of each trial, the RMS vector error and RMS error on the

Critical Task (if performed) were computed. This average RMS error score is used to determine the impact of the loading task on the primary tracking task.

Two further analysis techniques were employed using the tracking data:

1) Ensemble averages of display and control variables: In the Introduction section, it was argued than an important difference between detection performance during manual and autopilot control might relate to the relative importance of proprioceptive vs visual channels of information that the subject monitors as a basis for failure detection decisions. One means of proceeding in the analysis of relative signal importance is to sort the physical dimensions of the tracking data that follow each failure into categories defined by whether subjects did, or did not, detect the failure on that trial. If a particular dimension is found to differ between the two categories, evidence is provided that this variable was of use to the subject in his decision. That is, it represented a strong internal "signal" used by the subject to indicate failure occurrence. In the absence of this signal, failures were not detected.

To accomplish this sorting procedure, a technique of ensemble averaging was applied. Samples of the absolute tracking error, absolute control velocity and absolute cursor velocity were recorded at intervals of 60 msec. Ensemble averages time-locked to failure occurrence were then computed across all failures within a given condition for each subject with separate averages generated for detected failures (hits) and for misses. Naturally the control velocity measure was only averaged in the manual mode. The output of this analysis then was a series of average profiles of error and stick response to failures

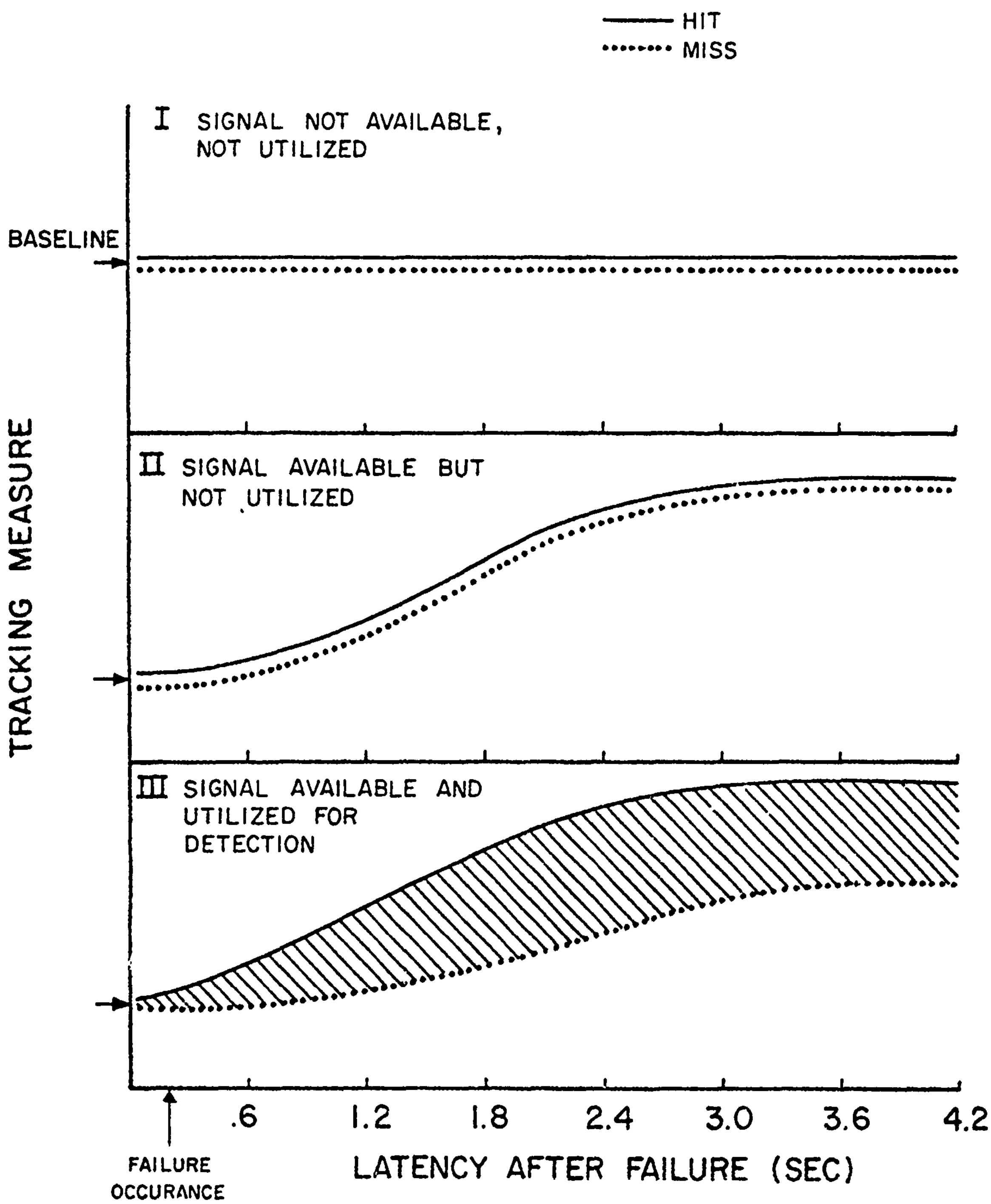


Figure 5. Idealized prototypes of tracking ensemble averages following failure.

(Young, 1969) that describe the time-course of these signals on trials when a fialure was detected and on trials when it was not.

Three ideal prototypes of these profiles are presented in Figure 5. The horizontal arrow on the ordinate indicates the prefailure baseline measure of the plotted variable. The vertical arrow represents the beginning of the failure while the shaded area to the right of the arrow represents the degree to which the plotted variable on hit conditions exceeded that on the miss conditions. A profile similar to Type 1 indicates that the physical signal did not change in response to the failure, and, therefore, that that information could not be employed as a basis for detection.

The extent to which either the hit or miss profiles (or both) rise above the pre-stimulus baseline (Type II and Type III) reflects the extent of usable information inherent in the physical signal indicating that a failure has occurred. To the extent that this physical information is actually <u>used</u> by the subject to make his decision concerning signal presence, the hit and miss profiles should be separated (Type III).

Ensemble averages time locked to the trigger press were also computed, thereby producing separate ensembles for hits as opposed to false alarms. This latter technique enables a more detailed comparison of the types of cues to which the subject is responding during the detection process, since cues that were in fact employed in detection should show overlapping profiles for hits and false alarms.

2) Multiple regression: The ensemble analysis of response and error profiles was designed to reveal global differences between hit and miss conditions under the assumption that signal characteristics that were relied upon to detect failures would be differentiated between

these two traces. A second approach taken to examine the cues utilized in detection was through a multiple regression analysis that determined which characteristic of the signals was most successful in predicting response latency on hit trials. If a cue used, coincided with a particular latency then its value should correlate strongly with response latency.

Values of the error, error velocity, control velocity, cursor velocity and Critical Task error (the latter under dual-task conditions), sampled at six time points following each detected failure, were employed as predictor variables of the criterion variable, detection latency. The time points sampled were at the instance of failure and at post-failure latencies of 0.6, 1.2, 2.4, 3.6, and 4.8 seconds. A stepwise multiple regression program (Biomed 02R) was employed in the analysis in which the data of the six subjects were collapsed within each experimental condition.

c) Statistical Power:

An analysis of the power of an experimental manipulation usually involves some post hoc technique to find the power of a test after the experiment has been performed. Cohen (1969) however, has argued in favor of employing a different technique to calculate the potential power of an experimental procedure before undertaking on the experiment. Cohen's procedure involves four main considerations, the alpha level one wants to chose to test the null hypothesis (i.e., the risk one is prepared to take in making a type I error); the number of experimental groups to be tested in a 1 way ANOVA; the number of subjects per experimental group and finally omega square or the proportion of variance that is expected due to the experimental manipulation. If all of the above are known one can calculate the probability of rejecting

the null hypothesis. This was done for the reported study: with an alpha level of .05, two independent experimental groups, 6 subjects in each cell and an omega square of at least .25 (a level chosen to be roughly consistent with the results obtained in the previous experiment, Wickens & Kessel, 1977) the null hypothesis was calculated to be rejected 90%[3] of the time. [Cohen considers a level of 70% to be acceptable]. A more conservative estimate of omega square, one equal to .20 predicted the rejection of the null hypothesis 80% of the time.

Results and Discussion

The statistical analysis for all the experimental comparisons was performed with a mixed model ANOVA (see Table 5) based on a Groups x Task Loading x Repetition (2x3x2) design.

There were two levels of Repetition--day 1 and day 2; 3 levels of the task loading factors--single task, dual easy and dual difficult and 2 groups were compared in each ANOVA. A separate ANOVA[4] was run for each of the comparisons specified in Figure 2:

(1) $MA_{I} - AU_{I}$

(2) $AU_{I} - AU_{II}$

(3) $MA_{I} - MA_{II}$

(4) $AU_{II} - AU_{II(c)}$

(5) $MA_{II} - AU_{II}$

(6) $MA_{II} - AU_{II(c)}$

(7) $AU_{I(c)} - AU_{II(c)}$ - within subject comparison

(8) $AU_{I} - AU_{I(c)}$

A separate ANOVA was run for each of the dependent variables:

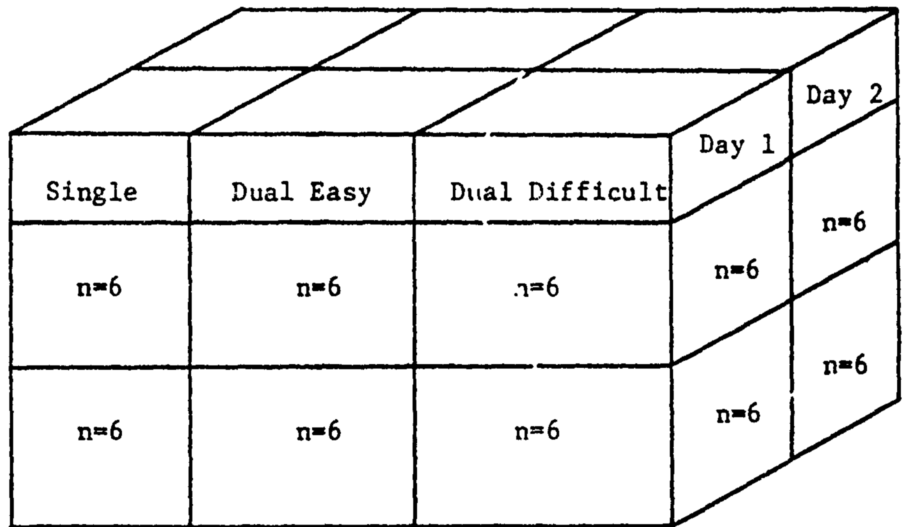
(a) Derived performance score

(b) P(A)

(c) P(HIT)

GROUPS

(Mode of Participation*)



REPETITION

TASK LOADING

NOTE: (1) The Mode of participation is a between subject manipulation

(2) Task loading and repetition factors were manipulated in subjects

---

*For experiment 3, this variable refers to the level of proprioceptive feedback i.e., MA with normal level and MA with reduced level.

Table 5

Mixed Model Three Way Analysis of Variance

(d) P(F/A)

(e) Latency

(f) Tracking performance - Absolute Error

(g) Critical task error - for dual task conditions only (here a 2x2x2 ANOVA was run).

Throughout the whole results section these ANOVA results will be reported for the reliable effects of the derived performance score only. Unless otherwise stated the main effects and interactions reported are those based on this 2x3x2 ANOVA analysis. All results pertaining to the dual task effect are discussed in greater detail in Wickens and Kessel (1979).

(A) Equality of Experimental Groups

All three groups produced virtually identical R.M.S. error scores on the pretest with the high levels of λ (see p. 19). Group 1 had a mean of .37 and standard deviation of .09; group 2 had a mean of .37 and standard deviation of .09; and group 3 had a mean of .36 and a standard deviation of .14. To the extent that this is a valid measure of tracking ability the three groups can be considered reasonably equal at the outset of the experiment.

(B) Detection

Averages and standard deviations were computed for the accuracy P(A), the latency and the derived performance measures following the rationale and the procedures outlined in the preceding section. These values are presented in Tables 6, 7 and 8 as a function of both the experimental condition and workload level.

The group averages for all three measures are presented graphically in Figure 6 which represents the results for the single task condition.

		Session 1			Session 2		
		Manual (MA_I)			Automatic (AU_{II})		
		Single	CT_1^*	CT_2^{**}	Single	CT_1	CT_2
Group-1	$\bar{x}$	0.90	0.86	0.86	0.90	0.87	0.87
	σ	0.03	0.05	0.04	0.04	0.08	0.04
		Automatic (AU_I)			Manual (MA_{II})		
		Single	CT_1	CT_2	Single	CT_1	CT_2
Group-2	$\bar{x}$	0.84	0.82	0.83	0.93	0.91	0.91
	σ	0.05	0.05	0.04	0.04	0.06	0.05
		Automatic (AU_{I-c})			Automatic (AU_{II-c})		
		Single	CT_1	CT_2	Single	CT_1	CT_2
Group-3	$\bar{x}$	0.84	0.86	0.85	0.89	0.86	0.86
	σ	0.04	0.06	0.06	0.03	0.04	0.05

* CT_1 = Easy Critical Task (λ = 0.5)
**CT_2 = Difficult Critical Task (λ = 1.0)

Table 6

Mean and Deviation Values for Accuracy P(A) As a Function of Experimental Condition And Workload Level

		Session 1			Session 2		
		Manual (MA_I)			Automatic (AU_{II})		
		Single	CT_1^*	CT_2^{**}	Single	CT_1	CT_2
Group 1	$\bar{x}$	2.3	2.7	2.8	3.1	3.4	3.4
	σ	0.4	0.5	0.6	0.6	0.8	0.6
		Automatic (AU_I)			Manual (MA_{II})		
		Single	CT_1	CT_2	Single	CT_1	CT_2
Group 1	$\bar{x}$	3.6	3.5	3.6	2.2	2.4	2.6
	σ	0.5	0.5	0.3	0.8	0.8	0.5
		Automatic (AU_{I-c})			Automatic (AU_{II-c})		
		Single	CT_1	CT_2	Single	CT_1	CT_2
Group 3	$\bar{x}$	3.3	3.6	3.6	3.4	3.6	3.5
	σ	0.5	0.3	0.5	0.5	0.3	0.2

* CT_1 = Easy Critical Task (λ = 0.5)
**CT_2 = Difficult Critical Task (λ = 1.0)

Table 7

Mean and Deviation Values for Latency as a Function of Experimental Condition and Workload Level (Seconds)

		Session 1			Session 2		
		Manual (MA_I)			Automatic (AU_{II})		
		Single	CT_1^*	CT_2^{**}	Single	CT_1	CT_2
Group 1	$\bar{x}$	6.7	6.03	5.85	5.9	5.3	5.3
	σ	0.29	0.88	0.73	0.9	1.53	0.89
		Automatic (AU_I)			Manual (MA_{II})		
		Single	CT_1	CT_2	Single	CT_1	CT_2
Group 2	$\bar{x}$	4.7	4.7	4.7	7.2	6.7	6.7
	σ	0.49	0.85	0.58	0.93	0.52	0.73
		Automatic ($AU_{I(c)}$)			Automatic ($AU_{II(c)}$)		
		Single	CT_1	CT_2	Single	CT_1	CT_2
Group 3	$\bar{x}$	5.1	5.0	4.86	5.5	5.02	5.1
	σ	0.83	0.70	0.98	0.74	0.59	0.52

* CT_1 = Easy Critical Task (λ = 0.5)
**CT_2 = Difficult Critical Task (λ = 1.0)

Table 8

Mean and Deviation Values for the Derived Performance Scores as a Joint Function of Experimental Condition and Workload Level

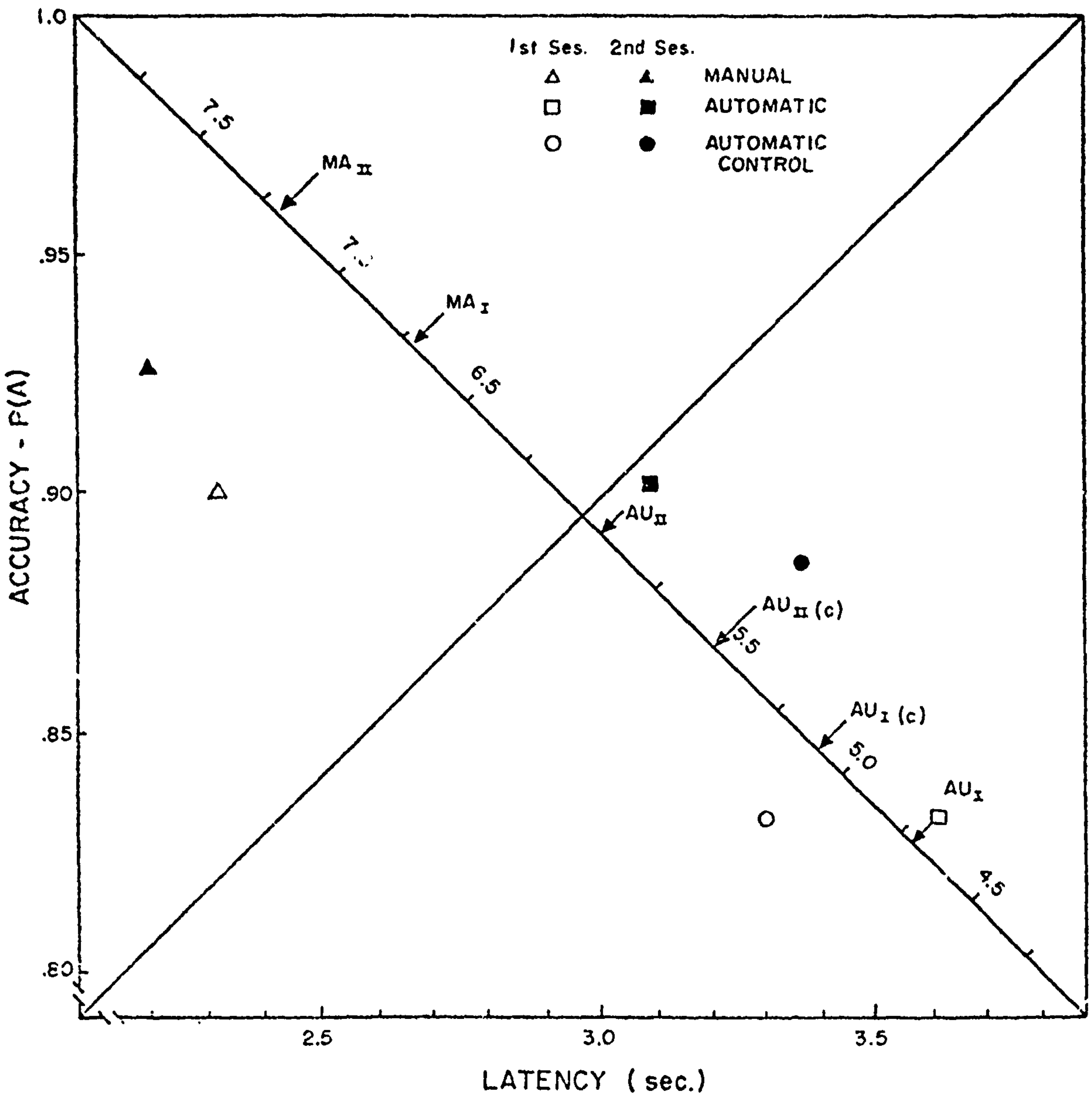


Figure 6. Effect of participatory mode and experimental condition on detection performance - Single Task.

The symbols in Figure 6 represent the group results in the speed-accuracy space, while the arrows and labels depict the derived performance scores for the various groups along the performance axis. In Figures 7 through 10 the experimental groups are plotted with the average derived performance score on the Y-axis.

The presentation of the results of the detection of failures will be divided into two sections. The first presents the results for each mode of participation, and represents a replication of the Wickens and Kessel (1977) study, while the second examines the results of the transfer of training experiment.

a) Mode of Participation

The most pronounced effect in the experimental data is the consistent superiority of MA over AU detection. This statistically reliable effect is clearly evident in the derived performance score shown in Table 8 and Figure 8 and was tested by contrasting group AU_I with MA_I ($F_{1,\ 10} = 18.4$, $p < .001$). Examination of Tables 6 and 7 and Figure 6 reveals that this superiority is reflected in overall detection latency ($F_{1,10} = 15.55$, $p < .01$), as well as accuracy ($F_{1,10} = 13.66$, $p < .01$).

While these findings essentially replicate the Wickens and Kessel (1977) study, it is important to note that the extent of MA superiority observed in the present results is greatly enhanced. In fact the magnitude of the MA-AU difference in the derived performance score is roughly five times its value obtained in the previous within-subject design. Contrasting the two studies, one finds that AU performance is unchanged, but MA performance in the present results is reliably superior to its level in the previous study ($t_9 = 2.18$, $p < .05$). This result suggests that in the previous experiment the AU internal model

was developed unhindered by the concurrent development of the MA internal model while the reverse situation did not hold.

It would appear that the development of the MA internal model in the previous experiment was somehow subject to interference from the AU model development, suggesting that subjects were paying attention to non-relevant, visual cues. As noted in the Introduction the sensitivity to proprioceptive information is reduced relative to visual information particularly when the two sources are available at the same time and are conveying conflicting information (see Posner et al., 1976; Klein & Posner, 1974; Jordan, 1972; Adams et al., 1977). In the AU mode the subjects have only visual cues as information while in the MA mode both visual and proprioceptive information are available. Thus during the development of the MA internal models there were times when these cues might be in conflict and subjects tended to fall back on the visual cues learned in the AU mode. This produced an over-emphasis on the visual cues and a subsequent degrading of the proprioceptive information. The introduction of the between subject design in the current experiment forced subjects to develop separate internal models based upon the relevant cues available within each condition--a situation that has enhanced the MA-AU differences found in the previous experiment.

By comparing the single task performance in MA_{II} with AU_{II} (see Table 8 and Figures 6 and 7) it is possible to determine whether MA superiority is maintained after prior training in the other mode of participation. From Figures 6 and 7 we can see that while this difference has been reduced somewhat the overall MA superiority remains intact. This MA_{II} - AU_{II} group difference is also statistically reliable ($F_{1,10} = 6.76$, $p < .05$), though from the figures it is clear that these MA_{II} - AU_{II} differences are somewhat smaller than those for

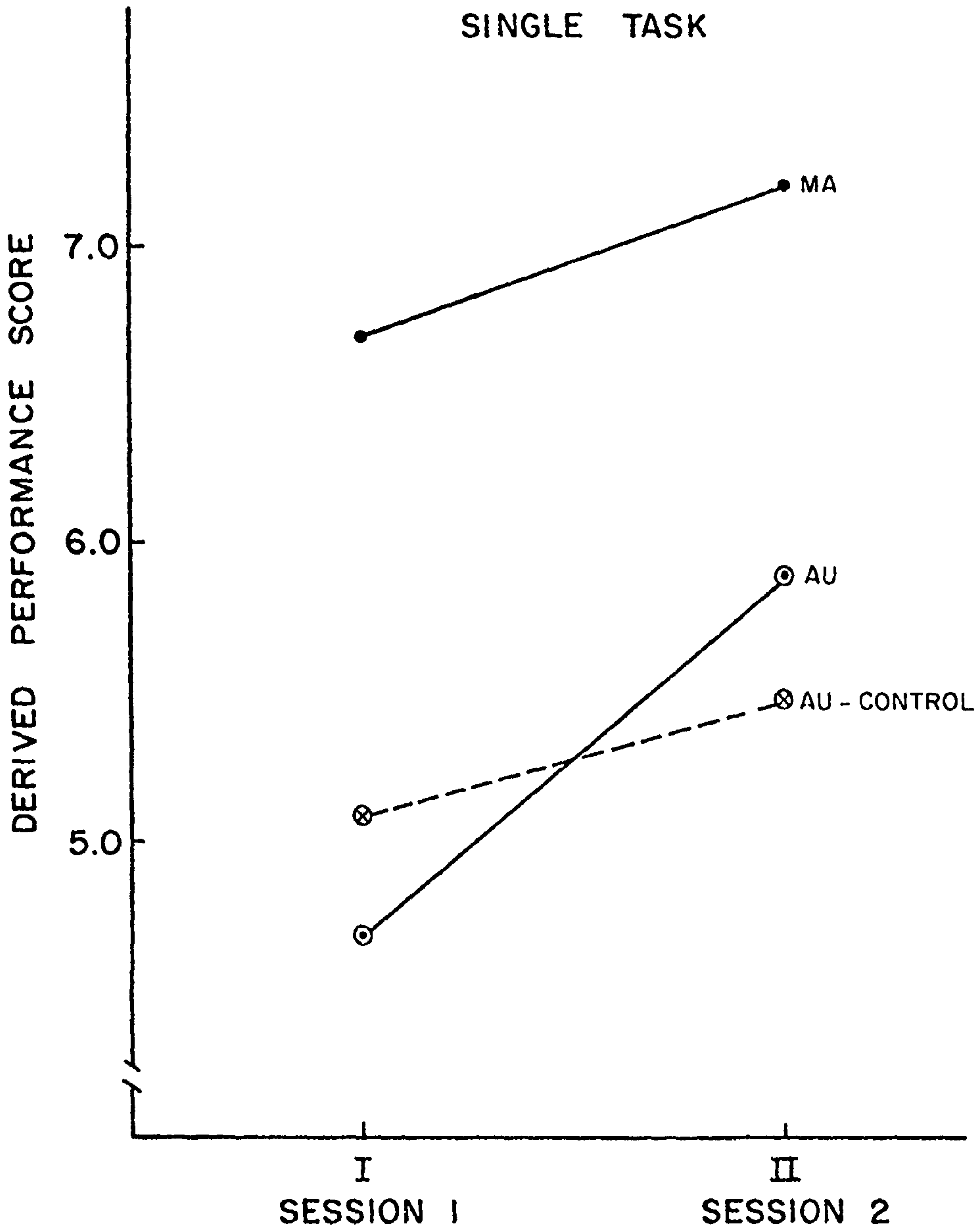


Figure 7. Detection performance as a function of experimental condition - Single Task.

the MA_I - AU_I comparisons.

These findings add strength to the argument that internal models developed separately tend to be more consistent, less variable and more sensitive to system changes.

b) Transfer of Training

Manual mode. In determining the relative amount of transfer to the manual mode resulting from prior automatic training, the MA_{II} group is compared with its control group MA_I (Figure 2) which essentially had no prior experience in the failure detection task.

From Table 8 and Figures 6 through 8 it can be seen that in general there is an overall MA_{II} superiority over MA_I for both single and dual task conditions. However the ANOVA failed to reveal these differences to be statistically reliable.

When the data are plotted on a day by day basis (see Figure 9) it is clear that any overall MA_I - MA_{II} difference is due mainly to the large differences that exist on day 1 which appear to dissipate completely when the two groups are compared on day 2 performance.

It can be concluded therefore that the MA_{II} group does not appear to benefit greatly from prior AU_I training. While this trend holds firm across all the experimental conditions the reasons for this apparent lack of transfer are not clear and could be either due to the greater experience of the MA_{II} group in the overall experimental situation or the fact that some transfer did take place (as indicated by the large day 1 group differences) but dissipated as the MA_{II} group rapidly developed detection strategies based on an internal model that relies more heavily on the extra information channel now available.

This finding substantiates the argument that the development of the internal model during the manual mode cannot utilize to advantage the

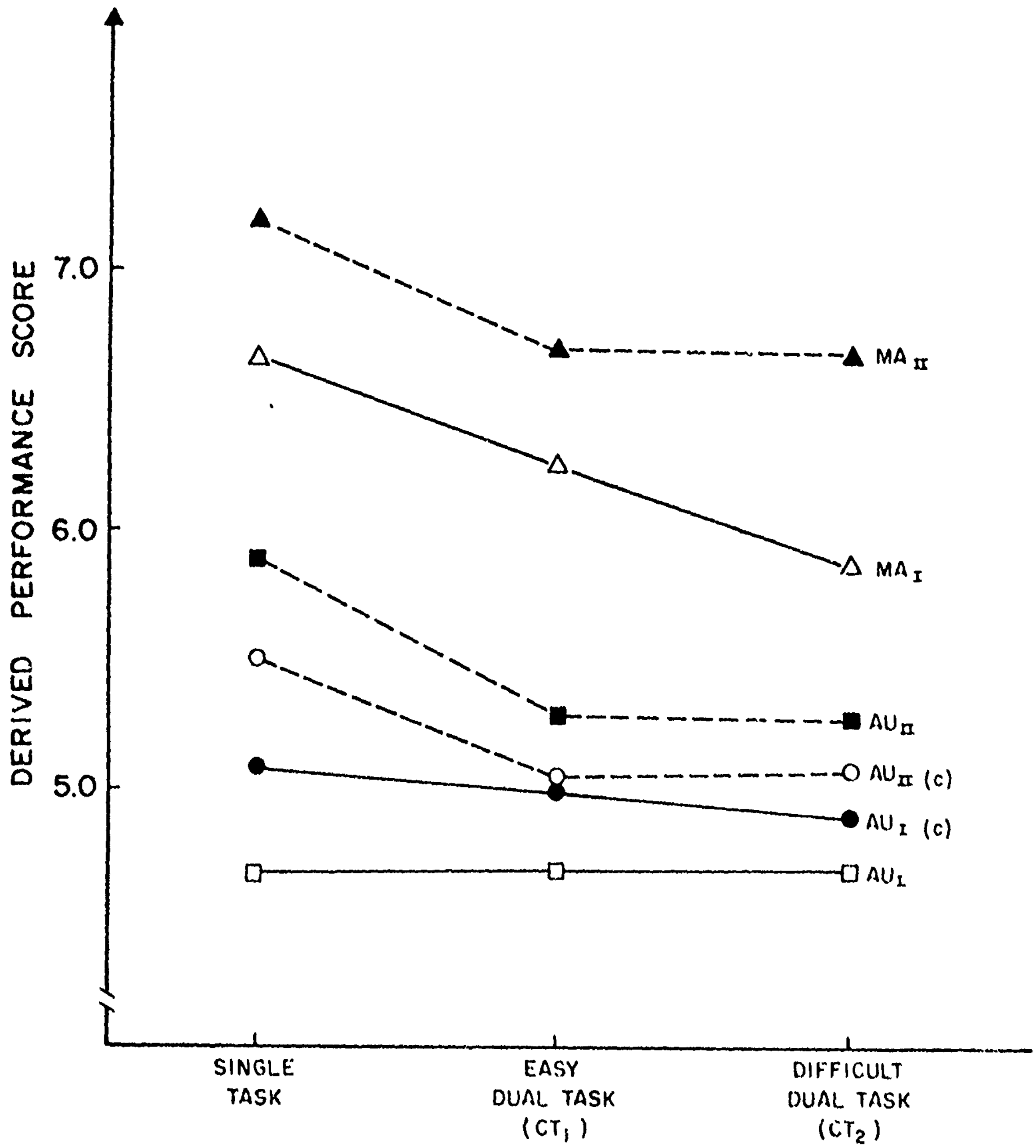


Figure 8. Detection performance as a function of concurrent task workload.

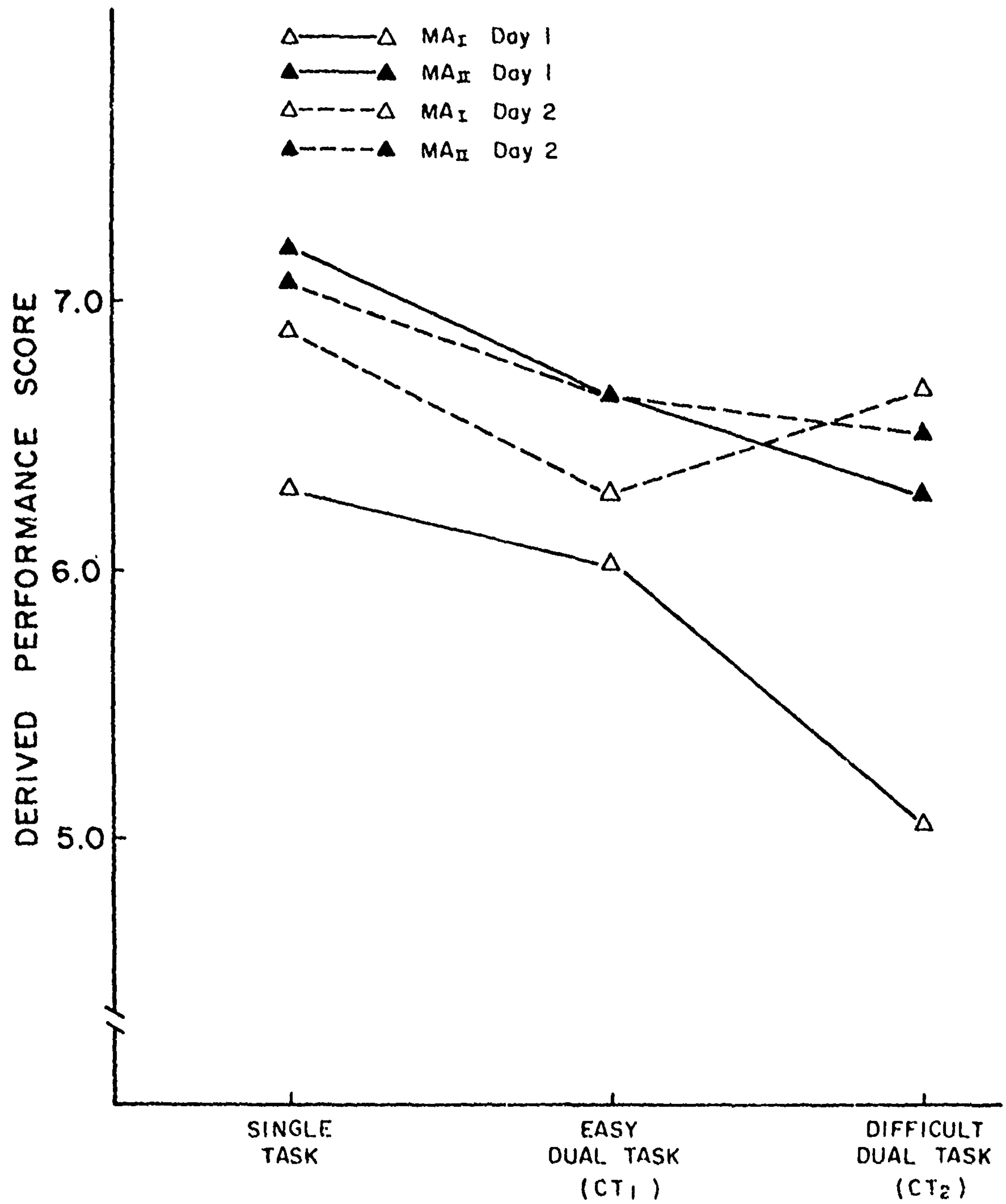


Figure 9. Detection performance in the manual mode as a function of concurrent task workload - by days.

internal model developed during the automatic mode. The addition of the proprioceptive channels and the interactive describing function in the manual mode appears to require the development of a separate and unique internal model.

Automatic mode. The degree of transfer resulting from prior MA training to the AU mode is reflected in the performance of subjects in condition AU_{II}, and the comparison of this performance with that of the control group ($AU_{I(c)}$ - $AU_{II(c)}$). In Table 8 and Figure 7, it is evident that the latter group failed to benefit at all from prior AU training, an observation supported by the lack of statistical reliability of the main effect when $AU_{I(c)}$ and $AU_{II(c)}$ are compared.

In marked contrast from Table 8 and Figures 6-8 it can be seen that the AU_{II} group in fact showed considerable benefit from their prior MA training when their performance is contrasted with that of the AU_I group. In Figure 7, the magnitude of this effect is seen to be considerably larger than the effect for the control group or for the MA_I - MA_{II} contrast discussed in the preceding section. The statistical reliability of this improvement was assessed by a groups (AU_I vs. AU_{II}) x days (Day 1 vs. Day 2), 2x2 ANOVA for single task condition only.

Both main effects were statistically reliable. This indicates that (a) both groups improved with practice (over two days) in their respective AU conditions ($F_{1,10} = 14.77$, $p < .001$). (b) More crucially, from the viewpoint of the hypothesis under investigation, the AU_{II} group performed reliably better than did the AU group ($F_{1,10} = 5.19$, $p < .05$). It is of course possible to argue that this effect resulted from greater exposure to and familiarity with the overall experimental environment experienced by the AU_{II} group and not to transfer of the internal model. However, this interpretation appears unlikely because

the control group failed to show any such "generalized" transfer.

When the data are examined on a day by day basis (Figure 10) it can be seen that this difference is largely due to a big improvement for the AU_{II} group on day two. This effect is more marked for the single task and dual task-easy conditions. This result tends to indicate that the prior development of an internal model with MA training facilitates failure detection performance on the second day of practice.

We can conclude that there is a transfer from MA to AU. The AU_I - AU_{II} differences are very large and statistically reliable and as such support the basic hypothesis that while there are different sets of cues operating, the MA condition produces an internal model of the system that can be utilized to advantage in subsequent automatic monitoring. Finally, these results tend to support the conclusion that the internal models developed in different modes of participation are relatively independent and therefore greater care must be exercised in extrapolating expected results in one mode of participation from performance in the other.

(C) Distribution of Detection Response Latencies

Figure 11 represents the frequency distribution of the response latencies for all the experimental groups. These results essentially replicate those reported in Wickens & Kessel (1977) in that the distributions for both the MA conditions were highly skewed in a positive direction, reflecting the shorter latencies. The AU groups on the other hand were approximately symmetrical with the noted exception of the AU_{II} group which has a distribution pattern that is similar to the MA groups for the first two seconds and then regains the AU pattern thereafter.

These latency distributions were transformed to cumulative

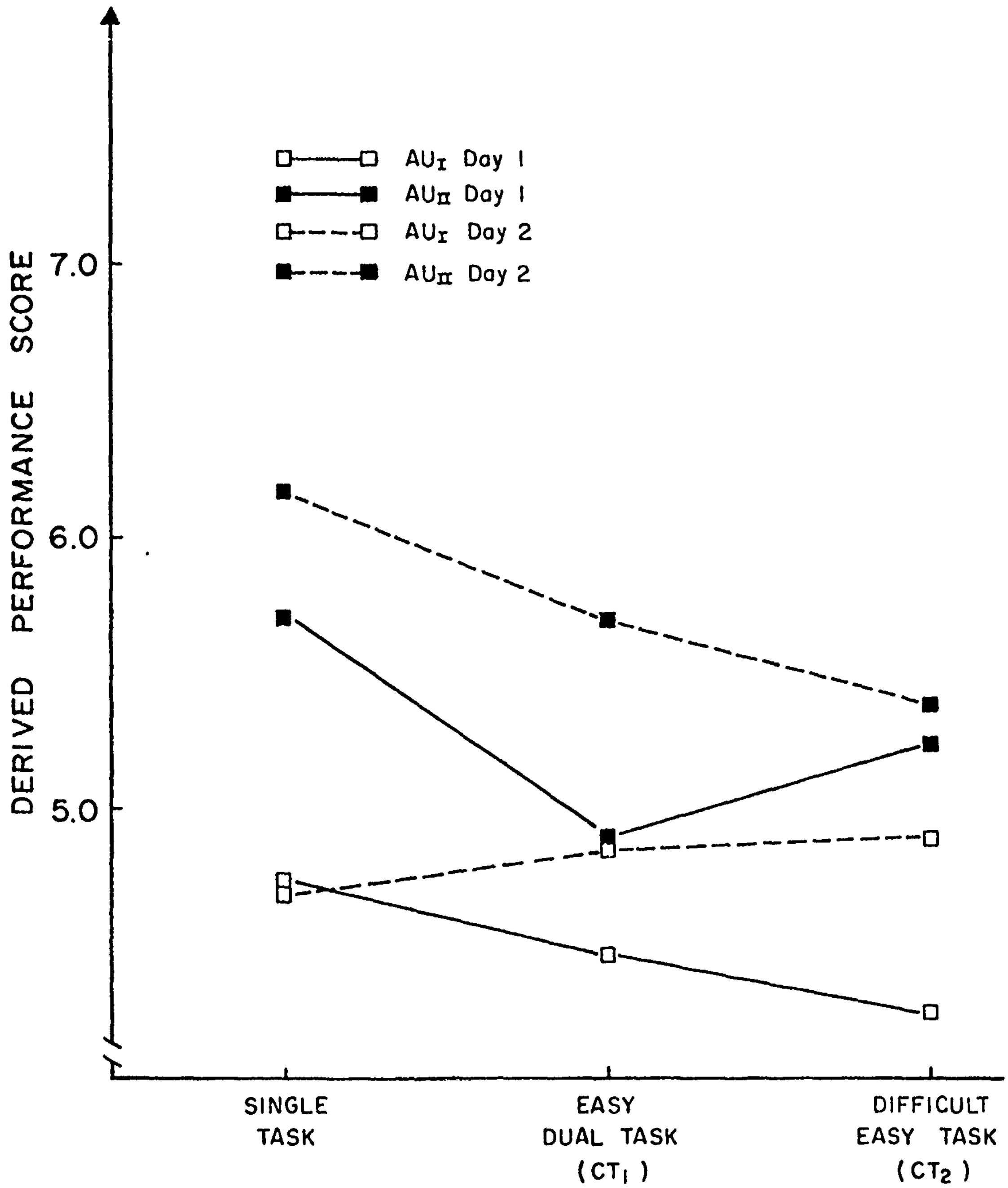


Figure 10. Detection performance in the automatic mode as a function of concurrent task workload - by days.

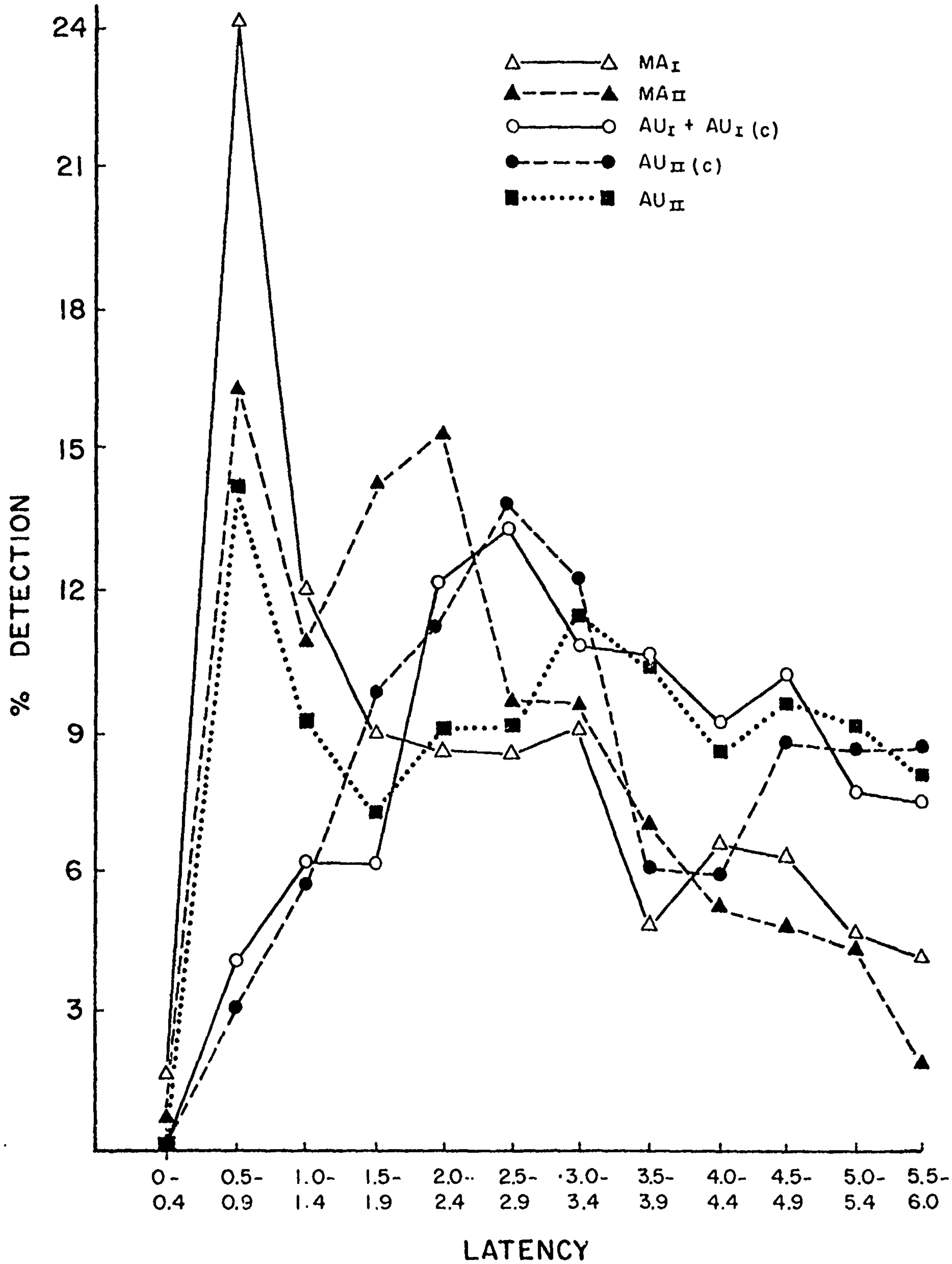


Figure 11. Latency distribution of detection responses following failure - by experimental condition.

probability distributions portraying the cumulative probability of failures detected, as a function of latency after failure (see Figure 12). These data once again show a clear replication of the Wickens and Kessel (1977) results with the noted exception of the AU_{II} group.

Following Lappin and Disch's approach (1972) (see analysis section above, p. 30) it can be argued that the slope of a function of relative accuracy vs. latency represents the rate at which perceptual evidence becomes available, while the level of the function or intercept represents overall quality of that information.

The interpretation of Figure 12 indicates that for the two MA functions there is a distinct discontinuity in the rate of accumulation of evidence, this discontinuity occurring at approximately 1-1.5 seconds post failure. The function describing AU_{II} performance while being slightly less marked appears to have a similar trend for both the single and dual task conditions. It would appear that the subjects in the AU group are able to take advantage of information available in their previous MA training condition to achieve this dichotomy (for clear evidence of this see AU_{II} in Figure 11).

The AU curves on the other hand while not strictly linear fail to slow the abrupt discontinuity of the MA conditions, and thus seemingly represent a uniform underlying process. In the Wickens and Kessel (1977) study the AU mode showed evidence of accumulating information at a faster rate (as represented by a steeper slope) than did the later, visual portion of the MA mode. No such difference was found in the present experiment and as can be seen from Figure 12 the slopes of the AU and MA modes are roughly parallel in the second part of the curve.

It would seem therefore that the development of separate internal models has provided information that is utilized very rapidly at the

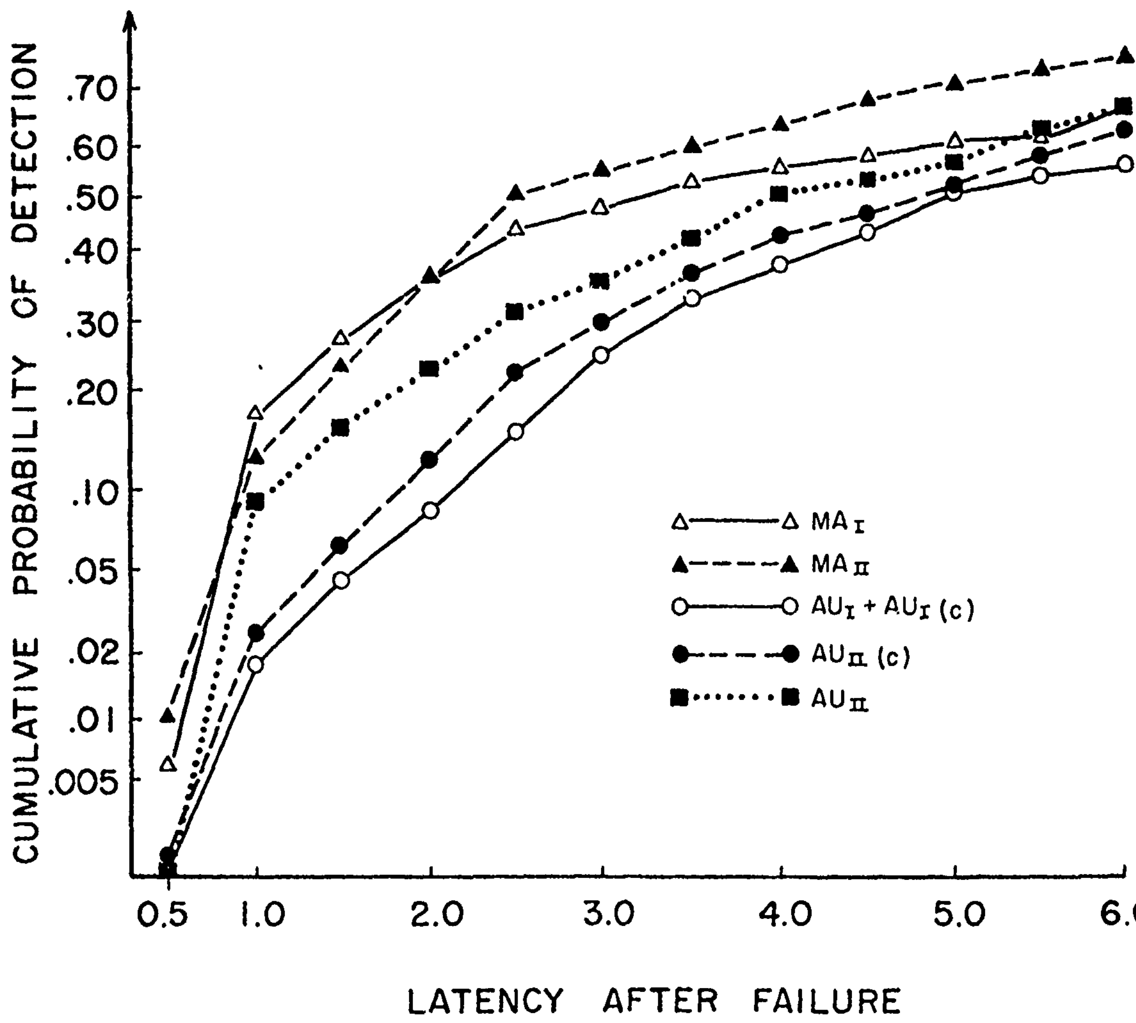


Figure 12. Cumulative probability distribution of detection latencies for single task automatic and manual mode conditions.

outset in the MA modes and the AU_{II} mode which is then integrated at roughly the same rate thereafter. The question that must be answered relates to the specificity of information available and utilized in the early stages by the AU_{II} group that were not available for either the AU or the $AU_{I(c)}$ and $AU_{II(c)}$ control groups.

Wickens and Kessel (1977) accounted for the MA-AU difference in the cumulative latency function by arguing that in the manual mode subjects utilize to advangate the additional proprioceptive channel. These AU_{II} findings tend to contradict this conclusion since no proprioceptive information is available in this condition, and seem to suggest that the manual control forces subjects to identify important visual cues normally ignored in the automatic mode. The ensemble average analysis below has addressed this particular question in an attempt to identify the specific cues used in the AU group that were not utilized by the other automatic groups.

(D) Ensemble Averages

As pointed out on page 32 above the ensemble averaging technique can be used to determine both the presence of cues and their utilization by the subjects in responding to system changes. By comparing the hit and miss profiles for the post failure ensembles and relating them to the hit and false alarm profiles for the pre-trigger ensembles, it is possible to arrive at some overall conclusion about the relative importance of the cues utilized in the decision making process. Furthermore, a comparison of the various experimental groups can reveal how different groups utilized different sets of available cues.

Ensemble averages were constructed of all the dependent variables listed on page 36 for both post failure and pre-trigger ensembles, and for all the experimental groups. Separate ensembles were plotted for

single and dual task conditions but since no meaningful differences emerged between these conditions only the single task ensembles are reported.

The three variables that proved to be the most illuminating are presented in Figure 13 (absolute error), Figure 14 (absolute control velocity) and Figure 15 (absolute cursor velocity). Each of these variables will be discussed separately.

a) Absolute Error

As can be seen in Figure 13, absolute error represents a clear example of Type III profile (see Figure 5) in that the error profile for all the traces in both the failure and trigger ensembles demonstrate a sharp rise from the average baseline condition. Furthermore the contrast between hits and misses is evident showing the strength of the error signal that is used for detection in both the MA and AU conditions. The lack of contrast between hits and false alarms, on the other hand, for the pre-trigger ensembles tends to add support to the argument that absolute error was an important variable in the decision making process for both MA and AU conditions.

A comparison of the $AU_I - MA_I$ profiles for the post failure ensembles reveals that the AU profile is higher for both hits and misses. This finding supports the conclusion that the MA tracker is showing a greater degree of adaptation than the AU tracker and is responding in such a manner as to attenuate his error. The greater separation between hit and miss profiles for AU than for MA supports the assumption that in the AU mode subjects relied more heavily on this cue than did the subjects in the MA mode. For example, at the 2.4 second latency AU_I hit-miss differences were twice as large as those in the MA_I group.

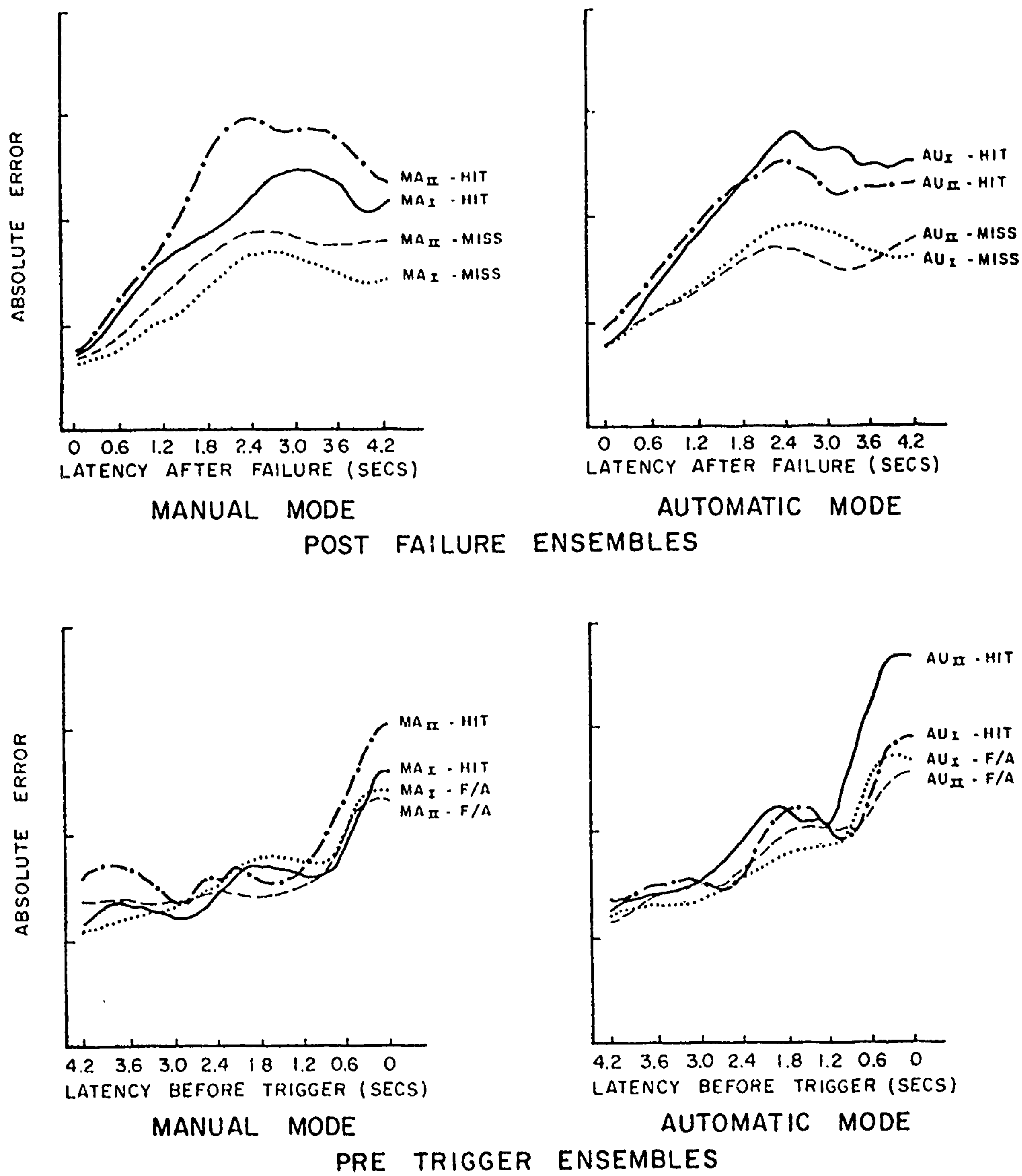


Figure 13. Ensemble averages of absolute error for automatic and manual conditions for post failure ensembles (top) and pre-trigger ensembles(bottom).

A further comparison between the different groups shows that on the whole the MA_{II} group adopted a strategy that produced greater absolute errors for both hits and misses than did the MA_I group. The existence of a large separation of hit from miss profiles following the 1.2 second interval indicates that the MA_{II} group developed a greater utilization of this cue than did the MA_I group. This finding is further corroborated by the differences between these two groups for the last 1.2 seconds prior to the trigger. In the trigger ensemble of the AU groups the AU_{II} group had a final 2 second profile that is different from the AU_I group. This result suggests that for this AU_{II} group the absolute error cues are playing a slightly different role than for the other groups.

b) Absolute Control Velocity

The ensembles of the absolute control velocity are presented in Figure 14. It should be recalled that since the AU tracking was computer generated no control velocity scores exist for this mode. The first salient result to emerge from the post failure ensembles is the large differences between hits and misses, indicating a fairly sharp utilization of this cue in responding to failures. The importance of control velocity is further supported by the trigger locked ensembles where for both the MA_I and MA_{II} groups there are no clear differences between hits and false alarms (see page 32 above).

An interesting result to emerge from both the failure and trigger ensembles is the relatively large difference in the hits profiles between MA_I and MA_{II}. This difference tends to reflect a differential control strategy adopted by the subjects in these groups and could well account for the increased absolute error produced by the MA_{II} discussed above, and its greater utilization as a cue.

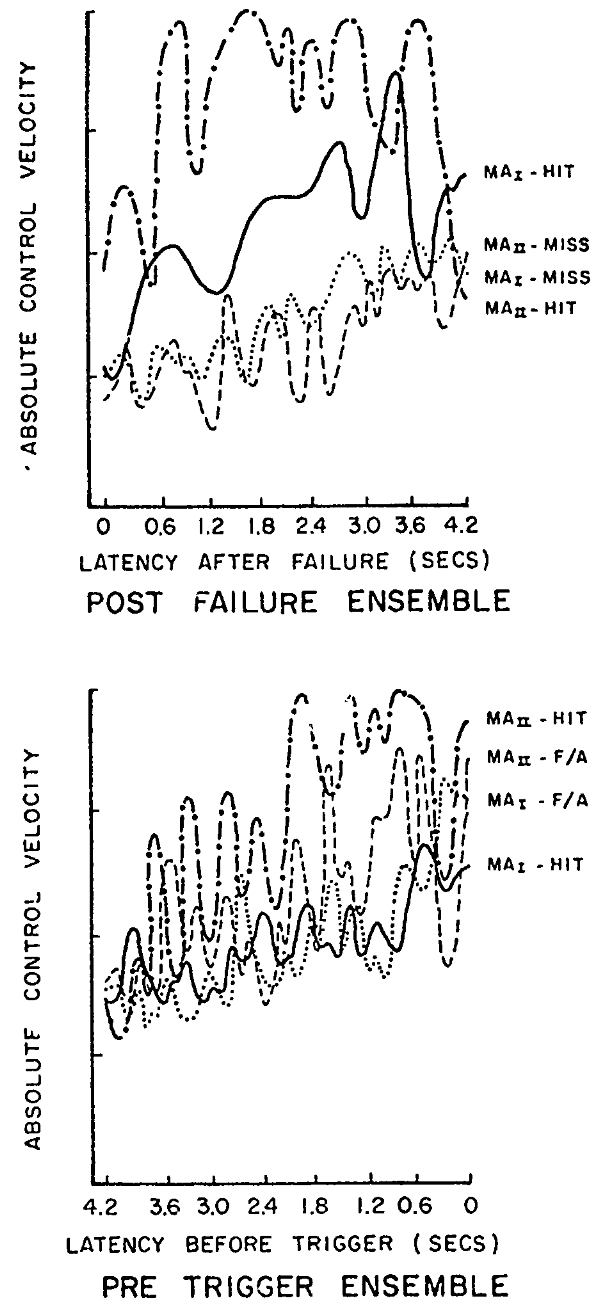


Figure 14. Ensemble averages of absolute cursor velocity for post failure ensembles (top) and pre-trigger ensembles (bottom).

This increase in control velocity can be accounted for in a number of ways. Hess (1978) has pointed out that subjects change response strategies when systems change from first order to second order systems. If subjects do make these changes they will be reflected in an overall increase in control velocity.

This increase in control velocity could, on the other hand, reflect a deliberate attempt on the part of the subjects to inject artificial signals into the system in order to test out the system response. The fact that there is a large difference between MA_I and MA_{II} would add credence to this argument since the MA_{II} subjects reported attempts at recreating the types of errors they had come to expect during the previous automatic phase. To do this they would have had to embrace a strategy of greater control movements. Such a strategy would account for both the absolute control velocity results as well as the fact that the MA_{II} had consistently greater absolute error profiles and greater tracking error scores (see Wickens & Kessel, 1979a) than did the MA_I group, and errors and hits were more separated in that group.

c) Absolute Cursor Velocity

From Figure 15 it can be seen that absolute cursor velocity ensemble profiles show clear cut differences between hits and misses for the failure locked ensembles while no differences between hits and false alarms for the trigger locked ensembles. These results clearly suggest that cursor velocity is an important cue utilized by all the manual and automatic groups.

A close scrutiny of the initial period of the post failure ensemble averages has revealed important differences between the groups. Both the MA groups show rapid sensitivity to the deceleration in the cursor velocity, a fact represented by the initial steep slope in the MA mode

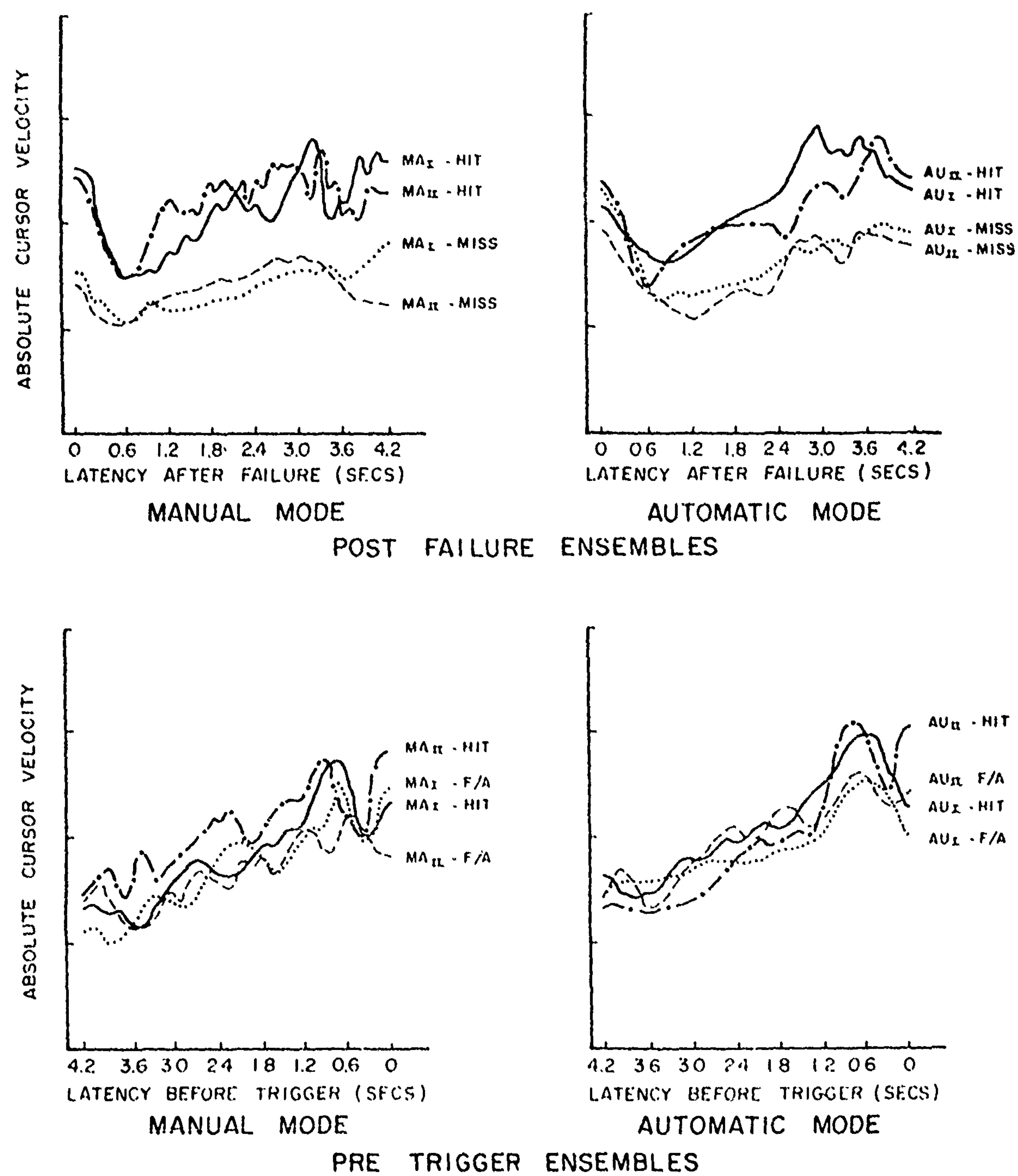


Figure 15. Ensemble averages of absolute cursor velocity for automatic and manual conditions for post failure ensembles (top) and pre-trigger ensembles (bottom).

hit profiles for the 0-0.6 second interval and the shallow slope of the miss profiles. This particular pattern is also evident for the AU_{II} group but not the AU_I group. The similarity between the AU_{II} and MA group ensembles suggests that this cue has been established as an important one during the MA phase of the experiment and is utilized to advantage during the subsequent AU_{II} phase.

In understanding these differences between the AU_{II} and other AU groups it is important to recall the nature of the system change. When a system changes from a velocity to an acceleration system, there is an initial deceleration of the target. Since the manual mode subjects have direct control of the cursor which they are using to follow the target, they develop a sensitivity to these changes in cursor velocity. They become increasingly sensitive to the sudden deceleration in the target when the system changes and also to its rapid acceleration thereafter. This information is therefore incorporated into their overall decision making process.

Subjects in the AU groups have no direct control over the cursor and therefore there is nothing that forces the subjects to take note of this particular dimension. However, the AU_{II} group having had the prior MA training, has the ability to utilize an internal model that has developed a sensitivity to the importance of velocity and acceleration factors.

(E) <u>Multiple Regression</u>

Employing the technique outlined above (see p. 32), a multiple regression analysis was run for each of the experimental conditions. The results of this analysis are presented in Table 9. The one variable that best predicts detection latency is presented together with the partial correlation. This partial correlation represents the square

	Session 1			Session 2		
	Manual - (MA_I)			Automatic (AU_{II})		
	Variable	Latency**	Partial r	Variable	Latency	Partial r
Single	Cursor Velocity	.06*	0.49	Error	1.2	0.32
CT_1	Error	1.2	0.32	Error	1.2	0.39
CT_2	Cursory Velocity	0.6	0.35	Error	1.2	0.36
	Automatic (AU_I+$AU_{I(c)}$)			Manual (MA_{II})		
	Variable	Latency	Partial r	Variable	Latency	Partial r
Single	Error	1.2	0.36	Cursor Velocity	0.6	0.33
CT_1	Error	1.2	0.39	Cursor Velocity	0.6	0.32
CT_2	Error	1.8	0.44	Cursor Velocity	1.2	0.21
				Automatic ($AU_{II(c)}$)		
				Variable	Latency	Partial r
Single				Error	0.6	0.46
CT_1				Error	1.8	0.35
CT_2				Error	1.8	0.35

* Seconds
** Latency of predictive variable following failure.

Table 9

Multiple Regression on Response Latency

root of the amount of variance accounted for by this variable. It should be pointed out that only one variable is presented in each condition since in all the experimental groups the second predictor variable accounted for only a very small amount of the additional explained variance.

The results in Table 9 show a clear and consistent dichotomy between the predictors for the Automatic groups and those for the Manual groups. It is interesting to note that only visually related variables (absolute error; and cursor velocity) proved to be predictors of response latency. For both the manual conditions, MA_I and MA_{II} , the dependent variable cursor velocity is best at predicting latency while for all the automatic groups the absolute error proved to be the best predictor.

These results are consistent with the ensemble averages in that cursor velocity did prove to be the most important cue for the Manual groups while absolute error was found to be the most discriminating cue for the Automatic groups. It is interesting to note that unlike the previous analysis, this multiple regression analysis did not produce differential results for the AU_{II} group. It should be pointed out however that while the ensemble averages analyzed accuracy of failure detection, this multiple regression analysis is concerned exclusively with predictors of response latencies.

EXPERIMENT 2

This experiment is discusssed in a separate Technical Report (Wickens & Kessel, 1979b, EPL-79-1/AFOSR-79-1). This report deals with the whole question of dual task performance and reports some additional experimentation.

EXPERIMENT 3

Subjects

The subjects were 6 male university students payed at the same rates and payoff schedules as in experiment 1.

Apparatus

This experiment employed the persuit tracking apparatus with an isotonic control stick. The control stick was the identical stick used in experiment 1 with the springs removed. The recorded resistence of the spring loaded control stick was 520 grams at maximum flexion. Since in the isotonic stick condition the springs were removed the resistence was zero.

Experimental Design

Since no side task was run the subjects only operated in the single task condition. The single task MA_I condition was therefore used as the control group for this experiment. This group, the proprioceptive group (P), received 3 sessions, one training session and two experimental sessions. Since only the single task condition was conducted subjects performed 10 trials on each of the experimental days and were exposed to 50 failures on each day.

Experimental Procedure

The experimental procedure employed by this group was identical to that for the MA_I single task condition (see experiment 1 above).

Data Collection and Analysis

The data collection and analysis procedures were identical to experiment 1 for all the detection and tracking performance scores. The

distribution of response latencies, ensemble averages and multiple regression analyses were also run on this group and compared to the MA control group. Since only the single task condition was run the mixed mode ANOVA employed a Groups x Repetition (2x2) design and was performed for the MA_I - P group comparisons for each of the dependent variables described in experiment 2.

Results and Discussion

a) Detection Performance

Averages and standard deviations were computed for the derived performance score, P(A), Latency and Absolute Error (for the rationale for each of these dependent measures see experiment 1, p. 36 above). These values are presented in Table 10 for both the proprioception and the MA control group.

The most striking result to emerge from Table 10 is the lack of any difference between the proprioception group and the MA group on all four of the dependent measures. The isotonic stick therefore did not degrade performance for either the detection or the tracking measures. This finding is in direct contrast with most of the experiments in the literature that have reported results with the isotonic stick. All these studies report clear superiority in tracking performance for the spring loaded stick over the isotonic stick (Gibbs & Baker, 1957; North & Lomnicki, 1961; Burke & Gibbs, 1965; Curry & Ephrath, 1976).

b) Tracking Analysis

The proprioception group had a response latency distribution that was almost identical to the MA_I group once again repeating the overall pattern for all the manual control groups. Like the MA_I single task condition the best predictor of response latency in the multiple regression analysis was cursor velocity (partial r = .29).

	Proprioception (P)		Control Group MA_I (Single Task)	
	$\bar{x}$	σ	$\bar{x}$	σ
Derived Performance Score	6.6	0.73	6.7	0.29
Accuracy P(A)	0.91	0.03	0.9	0.03
Latency	2.5	0.84	2.3	0.4
Absolute Error (ERR)	.12	0.03	0.11	0.03

Table 10

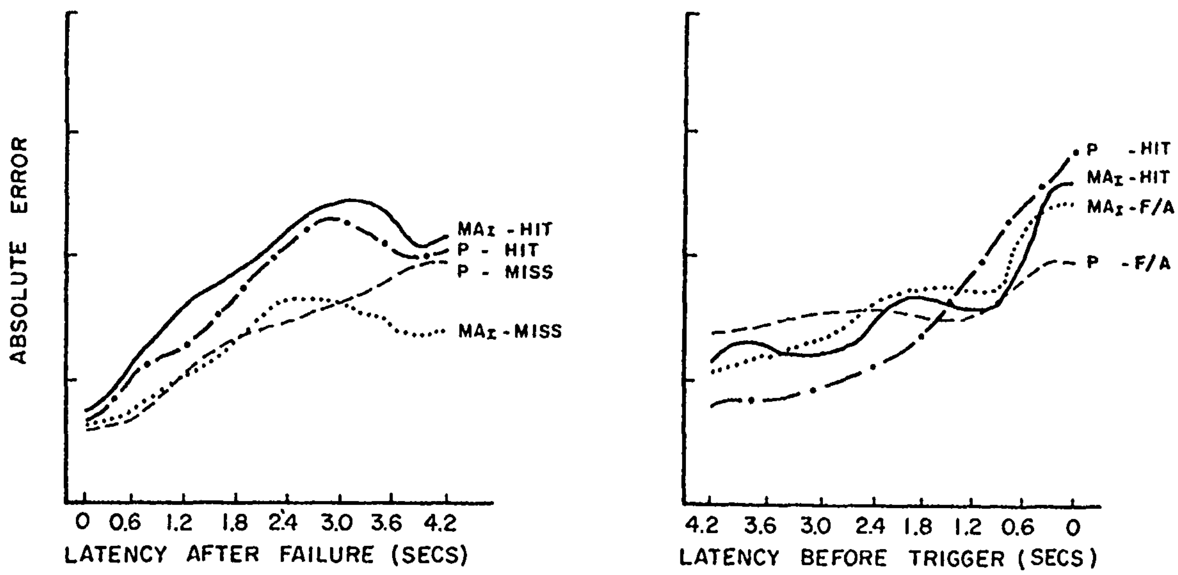
Mean and Deviation Values for the Proprioception Group (P) and Control Group (MA_I)

Some interesting differences between the proprioception group and MA emerged in the ensemble average profiles. Figure 16 represents the post failure and pre-trigger ensemble averages for the proprioception group together with the control group MA_I for the three dependent measures: absolute error (Figure 16a), absolute control velocity (Figure 16b) and absolute cursor velocity (Figure 16c).

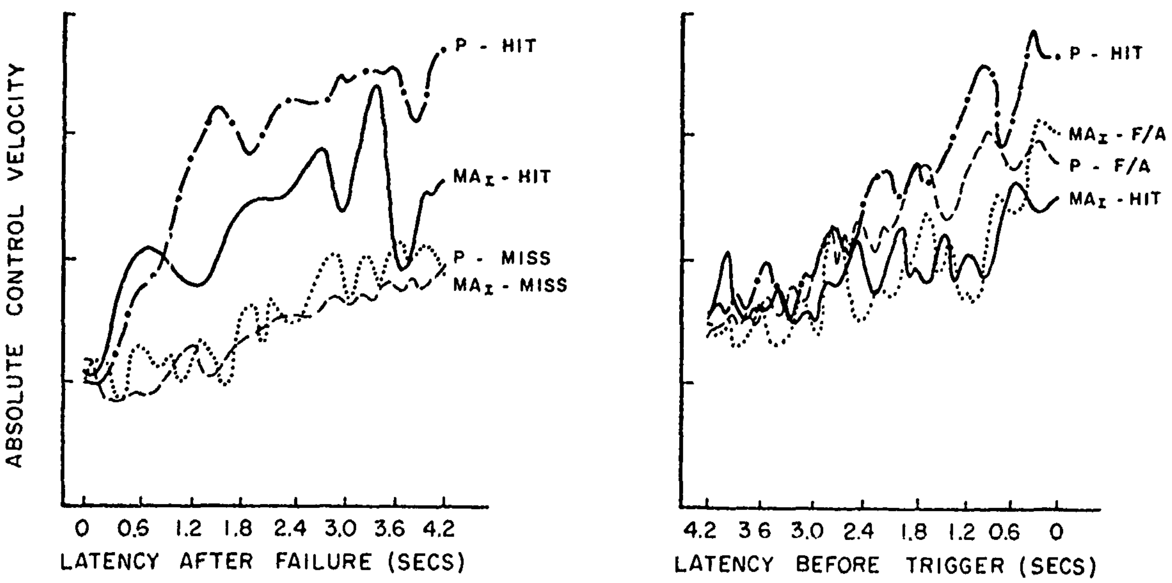
As can be seen in Figure 16a the MA_I group shows the classic type III profile (see p. 34 above) demonstrating the clear utilization of absolute error as a cue in discriminations between hits and misses. The proprioception group, on the other hand, has ensemble average profiles for both the post failure and pre trigger ensembles that demonstrate that this cue was of less value in discriminating between the hits and misses. This conclusion is reached from the comparison of the hits and miss profiles for the post failure analysis that are much closer together than the MA_I group while the hits and false alarm profiles in the pre-trigger ensembles were different.

The profiles of absolute control velocity show that the proprioception group had relatively larger hit profiles than the MA group. This difference tends to reflect a differential control strategy adopted by the subject or could conceivably be the result of the lighter control stick reflecting a greater number of reversals and movements of the stick. These control velocity ensembles are in many respects markedly similar to the profiles produced by the MA_{II} group in experiment 1.

Finally, from Figure 16c we can see that for the cursor velocity the proprioception group shows a greater sensitivity to the initial cursor deceleration than the MA_I group. This finding is illustrated in the initial 0.6 second post failure ensemble. It was argued in

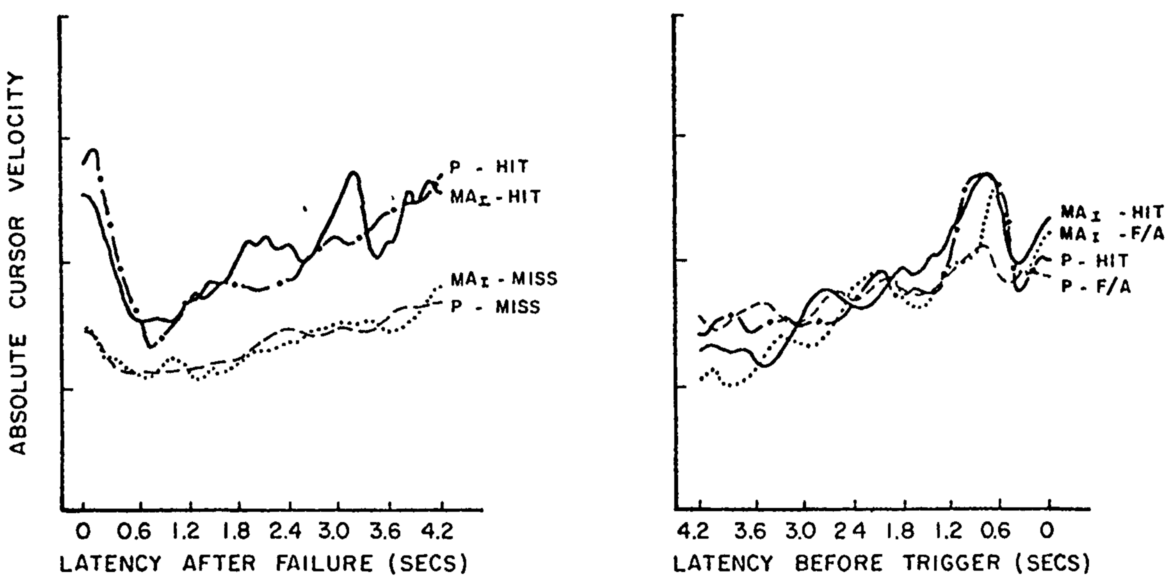


a) Absolute error.



b)Absolute control velocity.

Figure 16. Ensemble averages of manual tracking (MA_I) and isotonic stick (P.) conditions for post failure and pre-trigger ensembles.



c) Absolute cursor velocity.

experiment 1 above (p. 62) that sensitivity to cursor velocity changes is an important differentiating factor between manual and automatic modes of operation. What this finding tends to suggest is that this variable can also differentiate between different levels of sensitivity of the control stick and the value of this cue is apparently accentuated when the resistence of the control stick is reduced.

These ensemble results tend to suggest that while there are no differences betwen the groups in detection performance, the overall results have been reached by the adoption of different detection strategies. The proprioception group was relying less on absolute error cues while using the greater control sensitivity to pick up initial changes in cursor velocity. These results are consistent with the notion that the reduction in the resistence of the control stick has forced subjects to increase their overall sensitivity to system changes. While this sensitivity did not translate into higher failure detection rates it was reflected in a differential control strategy.

GENERAL DISCUSSION

From the results of the three experiments discussed above it is clear that the first two hypotheses (see p. 15) relating to the differences between the two modes of participation and the role of proprioception and visual cues in internal model development were fully confirmed. The second two hypotheses relating to the impact of workload and the strength of the proprioceptive feedback, on the other hand, were only partially substantiated. Some of the implications of these conclusions are discussed above. The aim of this section is to provide an overall theoretical framework for the findings of all three experiments and to suggest areas for further research. Some attention will be paid to the practical implications of these results.

The finding that the manual mode of operation was superior to the automatic in the failure detection task represents a replication of both Young (1969) and Wickens and Kessel (1977). This finding is interesting since unlike the previous research the present study adopted a between subject design which not only confirmed the manual mode superiority but appears to have increased its effect substantially. This can be considered strong support for the reliability of the conclusion that failure detection performance in the manual mode of operation is superior to performance in the automatic mode.

The results of the different experimental groups in the above three experiments add further weight to the overall reliability of this conclusion. It will be recalled that all the groups in all three experiments operating in the manual mode produced remarkably similar results both for overall detection performance and for the specific analyses of cue utilization. Likewise there was a strong reliability

effect for all the automatic groups in experiment 1 (with the noted exception of the AU_{II} group that benefited from prior manual experience). All these results suggest that the internal models developed in each of the experimental conditions appear to be both stable and reliable.

In the Introduction this superiority of the manual mode was attributed to the role of the proprioceptive channel, not available in the automatic mode. The transfer of training study was designed to specifically examine the role of this channel. It was also argued that this transfer technique would differentiate between the use of proprioception in the development and learning phase of the model as opposed to its role during the performance of the failure detection task. It was pointed out that the proprioception channel provides the subjects the opportunity to test out strategies in the failure detection task, though no direct evidence exists in any of the previous research to support this contention.

The first clue to the role of internal model development is provided by the comparison of the results of this study with those from the previous within subject design study (Wickens & Kessel, 1977). The between subject design produced manual superiority that was roughly five times as great as in the within subject design study. As noted above this difference can be attributed to the fact that the subjects were allowed to develop separate internal models for either the manual or the automatic mode, thereby producing models that were always appropriate for the mode of participation employed. This finding suggests that the way the internal model is developed is critical to its subsequent sensitivity in a failure detection task.

Further support for the importance of the mode of internal model

development was provided by the results in the transfer of training study. These results tend to show that all transfer subjects (i.e., both MA_{II} and AU_{II}) attempted to utilize the internal model developed in the previous mode of operation. The AU_{II} group obviously utilized the internal model developed during MA_{I} successfully. The MA_{II} group, on the other hand, experienced only brief and short lived benefits from the internal model developed during the AU_{I} condition. The MA_{II} superiority was retained during day 1 performance only.

The detailed analysis of the AU_{II} group revealed a remarkable similarity between characteristics of this group and all the MA groups. This similarity was no isolated phenomenon and is supported by the evidence from the overall detection performance, and from cue utilization provided by the cumulative frequency distributions, and the ensemble averages. On the whole the subjects in the AU_{II} group were better failure detectors than subjects in the other AU groups. Furthermore, the AU_{II} groups cumulative frequency distribution was very similar to the ones produced by all the other MA groups, showing a tendency for rapid accumulation of data not found in any of the other AU groups.

The AU_{II} group uniqueness was further witnessed in their use of information relating to the changes in cursor velocity. Unlike any of the other AU groups the AU_{II} group placed greater importance on this cue in the decision making process and in so doing produced similar ensemble averages to the ones obtained for all the MA groups. The cursor velocity ensembles for AU_{II} clearly differentiated between hit and miss profiles while they showed particular sensitivity to the initial deceleration of the cursor as the system changed. This pattern of sensitivity to initial deceleration of the cursor was evident in all the

MA groups ensembles while none of the other AU groups showed this particular cue utilization pattern. Indeed, all these results tend to converge to one overall conclusion--that the AU_{II} group has developed a detection strategy that is both markedly different from all the other AU groups and markedly similar to all the MA groups.

The AU_{II} group results are particularly interesting in their implications for the role of proprioception, and the importance of this cue in MA performance: despite the close similarity between the MA and AU_{II} groups in cue utilization the MA groups continued to demonstrate overall superior detection performance. This would indicate that the role of proprioceptive information during failure detection performance cannot be discounted, indeed this channel operates both in the utilization of hypothesis testing by way of injecting artificial signals into the system, as well as in the adaptation of the subjects to system changes.

In understanding how subjects adapt to system changes it should be recalled that during the manual mode, operators are tracking a mixed first order $(\frac{K}{S})$ system. When the system changes to second order $(\frac{K}{S^2})$ operators are required to generate lead (differentiate, or respond directly to velocity of error and cursor) as KS dynamics in order to maintain stable tracking performance (McRuer, 1974). The proprioception channel therefore does not, in itself provide the information, rather it is the describing function that operators develop and are forced to change when a failure occurs, that accounts for this sensitivity. However, the clear transfer of information acquired with the proprioception channel to a situation when that channel is no longer available suggests that the manual mode superiority cannot be wholly accounted for by the describing function. In brief therefore it can be

concluded that proprioception is critical in developing internal models with enhanced sensitivity but its role during detection performance is supplemented by other characteristics also available to AU_{II}.

The proprioception channel, therefore, is critical during the development of the internal model but once having established the relevance of specific cues, its role during the performance phase is less critical. As noted above the continued MA superiority over the AU group suggests that the additional role of the proprioception channel, i.e., its hypothesis testing role continues to be important during the performance phase. It would be interesting to provide the AU subjects with the ability to make discrete tests of the system--to determine just how important this hypothesis testing function is. It should be noted that most of the MA subjects when debriefed after the experiment alluded to the use of such a testing strategy.

In conclusion therefore the results appear to support the notion that the manual mode is superior to the automatic mode and that this superiority can be attributed to the role of proprioception during both the development of the internal model and its utilization in the failure detection task. This requirement to control the system acted as a means of directing attention of the subjects to relevant cues during the initial phases of the system change, for example, cursor velocity and acceleration. The monitors in the automatic mode do not have this particular attention focusing mechanism, a fact reflected by their poorer detection performance. Finally the fact that monitors can use an internal model developed during the manual mode to advantage shows that internal models are not necessarily mode specific. When the cue sensitivity of a model developed in one mode is relevant to performance in another, generalization can take place.

The specific role of the proprioception channel described above has been supported by the results of the proprioception group in experiment 3. Changing the resistence of the control stick did not influence, as was anticipated in the introduction, the overall detection performance of the subjects. Although no differences were found in failure detection between the proprioception group that used the isotonic stick and the control group with the spring loaded stick the results suggest that these groups have adopted different cue utilization strategies. The proprioception group relied less than the MA control group on the absolute error cue while placing greater emphasis on the cursor velocity cue, thus emphasizing even further the differences between MA and AU modes of participation.

What this result suggests is that by changing the sensitivity of the control stick subjects are forced to change the way they control the system and this in turn results in the development of different cue dependencies. It is as if the greater sensitivity of the control stick has heightened the directional focus of the subject producing a much greater reliance on the cursor velocity cue and a smaller dependence on the absolute error cue. Unfortunately, the manipulation used in experiment 3 does not allow a clearer understanding of just how this change in control stick sensitivity is affecting the inflow and outflow of information. It would be interesting to examine the transfer of subjects from this group to the automatic mode. Should the above line of logic hold true then this transfer group should show even more dependence on the cursor velocity cue and less dependence on absolute error than have subjects in the AU_{I1} group in experiment 1.

Training Implications

These results suggest that when training manual operators and

monitors of automated systems care must be exercised in developing stable and invariant internal models. One conclusion from the above is that internal models developed in different modes of participation are relatively independent and therefore care must be made in extrapolating expected results in one mode of participation from performance in the other. In particular good monitors do not necessarily become good controllers, while good controllers can function as good monitors.

The results of experiment 1 tend to suggest that while transferring from monitoring to controlling only a little is gained in terms of failure detection, however, there is a decrement in manual tracking performance. This suggests that monitors are not always capable of instantaneously taking over the role of manual operation without prior experience. In applied settings monitors of automated systems usually have some manual experience and it is assumed that monitors can take-over from automatic controllers (such as auto-pilots) without difficulty thus serving as back-ups in times of potential failure.

The results from this experiment suggest that not much can be gained from a monitor to manual transfer unless it can be demonstrated that the "attention-focusing" role of the proprioceptive channel can be achieved by some other means. For example, by an exact verbal description of what visual cues to look for. A prom'sing line of research in this area would be to determine whether the manual mode superiority can be eliminated or reduced by providing subjects with this relevant information during the development phase. Subjects in the manual mode will always retain their ability to test out strategies, however, the impact of this could be radically reduced.

Another logical extension of this study would be to measure the monitors ability to intervene in an ongoing system when failures are

detected. This could be studied as a function of the nature and stability of their current internal model and measures could be obtained of just how often monitors should receive practice in operating the system. In other words, updating their own internal model in the manual mode.

Indeed the above experiments suggest that the best way of developing an internal model for monitors is to first provide them maximum access to the system as operators before transferring them to the monitoring mode. It is expected that this strategy will have a positive payoff in that they will develop greater sensitivity to specific system cues otherwise not attended to in the monitoring mode. This latter strategy is obviously less cost effective than the one of transferring from automatic to manual.

Finally, it would appear from the results of this study that controllers continue to be better failure detectors than monitors. In systems with severe consequences for undetected failures, as in the case of automatic landing systems in commercial aircraft, this fact should be taken into consideration. Furthermore taking operators out of the loop and turning them into monitors does not automatically ensure their ability to operate as back-up systems in times of failure. Consideration should be given to interactive systems that will take advantage of the computer while not losing the failure detection ability of the controller. This could be achieved by requiring the operator to maintain a link with the system being run by the automatic controller, the auto-pilot. This will ensure that the operator retains a well-updated internal model of the dynamics of the system which will result in a greater sensitivity to system failures and the ability to successfully take over the control, should the need arise.

BIBLIOGRAPHY

Adams, J. A., Gopher, D., & Lintern, G. Effects of visual and proprioceptive feedback on motor learning. Journal of Motor Behavior, 1977, 9 (1), 11-22.

Buck, L. Reaction time as a measure of perceptual vigilance. Psychological Bulletin, 1966, 65, 291-308.

Burke, D., & Gibbs, C. B. A comparison of free moving and pressure levers in a positional control system. Ergonomics, 1965, 8, 23-29.

Cohen, J. Statistical Power Analysis for the Behavioral Sciences, New York: Academic Press, 1969.

Curry, R. E., & Ephrath, A. R. Monitoring and Control of Unreliable Systems. In T. B. Sheridan & G. Johannsen (Eds.), Monitoring Behavior and Supervisory Control, New York: Plenum Press, 1976.

Curry, R. E., & Gai, E. G. Detection of random process failures by human monitors. In T. B. Sheridan & G. Johannsen (Eds.), Monitoring Behavior and Supervisory Control, New York: Plenum Press, 1976.

Egan, J. P. Signal detection theory and ROC analysis. New York: Academic Press, 1975.

Egan, J. P., Greenberg, G. Z., & Schulman, A. I. Interval of time uncertainty in auditory signal detection. The Journal of the Acoustical Society of America, 1961a, 33 (8), 771-778.

Egan, J. P., Greenberg, G. Z., & Schulman, A. I. Memory for waveform and time uncertainty in auditory detection. The Journal of the Acoustical Society of America, 1961b, 33 (8), 779-781.

Egan, J. P., Greenberg, G. Z., & Schulman, A. I. Operating characteristics, signal detectability, and the measure of free response. The Journal of the Acoustical Society of America, 1961c, 33 (8), 993-1007.

Emerich, D. S., Gray, J. L., Watson, C. S., & Tanis, D. C. Response latency, confidence, and ROCS in auditory signal detection. Psychophysics, 1972, 11 (1B), 65-72.

Ephrath, A. R. Detection of system failures in multi-axes tasks. Proceedings of the 11th annual NASA-University conference on manual control. NASA TMX, 1975, 62, 464.

Ephrath, A. R., & Curry, R. E. Detection by pilots of system failures during instrument landings. IEEE Transactions on Systems Man and Cybernetics. Vol. SMC-7, No. 12, 1977, pp. 841-848.

Fitts, P. M. Cognitive aspects of information processing: III. Set for speed vs accuracy. Journal of Experimental Psychology, 1966, 71, 849-857.

Freedy, A., Davis, K. B., Steeb, R., Samet, M. G. & Gardiner, P. C. Adaptive computer aiding in dynamic decision processes: Methodology evaluation and applications. Report: ARPA, 1976.

Gibbs, C. G., & Baker, J. C. Free moving versus fixed control levers in a manual tracking task. In A. Tustin (Ed.), Automatic and Manual Control, London: Butterworths, 1957.

Green, D. M., & Swets, J. A. Signal detection theory and psychophysics, New York: John Wiley and Sons, Inc., 1966.

Hess, R. A. A rationale for human pulsive control behavior. Proceedings of the fourteenth annual NASA-University conference on manual control, NASA SP, 1978.

Jex, R., McDonnell, J. D., & Phatak, A. V. A "critical" tracking task for manual control research. IEEE Transactions on Human Factors in Electronics, 1966, 7 (4).

Jordan, T. C. Characteristics of visual and proprioceptive response times in the learning of a motor skill. Quarterly Journal of Experimental Psychology, 1972, 24, 536-543.

Klein, R. M., & Posner, M. I. Attention to visual and kinesthetic components of skills. Brain Research, 1974, 71, 401-411.

Kleinman, D. L., Baron, S., & Levison, W. H. A control theoretic approach to manned-vehicle systesm analysis. IEEE Transactions on Automatic Control, Vol. Ac-16, No. 6, 1971, 824-832.

Lappin, J., & Disch, K. The latency operating characteristic: I. Effects of stimulus probability. Journal of Experimental Psychology, 1972, 92, 419-427.

Luce, D. R. A model for detection in temporally unstructured experiments with poisson distribution of signal presentations. Journal of Mathematical Psychology, 1966, 3, 48-64.

McDonnell, J. D. A preliminary study of human operator behavior following step changes in the controlled element. IEEE Transactions on Human Factors in Electronics, HFE Tech. 1966, 125-129.

McRuer, D. T., & Krendel, E. S. Mathematical models of human pilot behavior: Report: NATO, AGARD-AG-188, 1974.

McNichol, D. A primer of signal detection theory. London: George Aver and Irwin Ltd., 1972.

North, J., & Lomnicki, A. Further experiments of human operators in compensatory tracking tasks. Ergonomics, 1961, 4, 339-353.

Pachella, R. The interpretation of reaction time in information-processing research. In B. H. Kantowitz, Human Information Processing: Tutorials in Performance and Cognitions, New York: Wiley, 1974.

Parasuraman, R., & Davies, D. R. Decision theory analysis of response latencies in vigilance. Journal of Experimental Psychology: Human Perception and Performance, 1976, 2 (4), 578-590.

Pew, R. W. The speed accuracy operating characteristic. In W. G. Koster, (Ed.), Attention and Performance II, Amsterdam: North Holland Publishing Company, 1969.

Pew, R. W. Human perceptual-motor performance. In B. H. Kantowitz, (Ed.), Human Information Processing: Tutorials in Performance and Cognition. New York: Wiley, 1974.

Pike, R. Response latency-signal detection. Psychological Review, 1973, 80, 53-68.

Posner, M. I., Nissen, M. J., & Klein, R. M. Visual Dominance: An information-processing account of its origins and significance. Psychological Review, 1976, Vol. 83, No. 2, 157-171.

Rouse, W. B. A theory of human decision making in stochastic estimation tasks. IEEE Transactions on Systems, Man and Cybernetics, Vol. SMC-7, No. 4, April 1977, 274-283.

Rouse, W. B. Design of man-computer interfaces for on-line interactive systems. Proceedings of the IEEE, Vol. 63, No. 6, June 1975, 847-857.

Sheridan, T. B., & Johannsen, G. Monitoring Behavior and Supervisory Control, New York: Plenum Press, 1976.

Swets, J. A., & Kristofferson, A. B. Attention. Annual Review of Psychology, 1970, 21, 339-366.

Veldhysen, W., & Stassen, H. G. The internal model: What does it mean in human control. In T. B. Sheridan & G. Johannsen, (Eds.), Monitoring Behavior and Supervisory Control, New York: Plenum Press, 1976.

Watson, D. S., & Nichols, T. L. Detectability of auditory signals presented without defined observation intervals. Journal of Acoustic Society of America, 1976, 59 (3), 655-668.

Wickens, C. D., & Kessel, C. J. The effects of participatory mode and task workload on the detection of dynamic system failures. Technical Report. ARL-77-8/AFOSR-77-7, June, 1977.

Wickens, C. D., & Kessel, C. J. The effects of participatory mode and task workload on the detection of dynamic system failures. IEEE Transactions on Systems Man & Cybernetics SMC-13, January, 1979a.

Wickens, C. D., & Kessel, C. J. The processing resource demands of failure decection in dynamic systems: Technical Report EPL-79-1/AFOSR-79-1, January, 1979b.

Young, L. R. On adaptive manual control. IEEE Transactions on Man-Machine Systems, V. MMS-10, 1969, 4, 292-331.

FOOTNOTES

1. The time delay is a parameter that is given in multiples of 60 ms. (the cycle time of the computer). This time delay therefore is based on 9 x 60 ms.

2. Single vector measures, of the form x + y are stored for both error and control position, rather than the separate x and y axis values because of tape and computer limitations.

3. This number is based on a set of unpublished power tables calculated by Allen Fleishman of the Department of Psychology, University of Illinois, 1977.

4. The fact that the same experimental group is being used in a number of different ANOVA's does affect the expected probabilities of establishing reliable results. This is, however, considered a legitimate procedure provided a priori reasons exist for these multiple ANOVA's as was the case in this experiment.

APPENDIX A

FAILURE DETECTION METHODOLOGY AND ANALYSIS

Performance in detecting system failures has conventionally been evaluated by assessing the proportion of detected (or missed) failures and/or the latency of detection. It is argued here, in support of analysis procedures derived by Curry and his colleagues (e.g., Curry & Gai, 1976) that added insight may be provided, first, by considering failure detection in the framework of statistical decision theory, and second, by integrating detection accuracy and latency measures as two interrelated indices of a common underlying process.

It is hypothesized that the operator/detector is consistently evaluating the current estimated state of the system and comparing it against an internal representation of its normal operating characteristics (Figure 1). The occurrence of a failure is an event which produces a change in information concerning the system state. This information is sampled and integrated by the operator over time and is compared with the representation stored in memory of "normal" behavior. Therefore, given that a failure occurs, the accuracy of the decision should increase with the integration time (reflected as decision latency). Finally the decision itself is governed by response bias--a decision criterion dictating the amount of disparate evidence that the operator considers sufficient to warrant a decision. In order to separate bias factors from sensitivity factors, the theory of signal detection (Green & Swets, 1966) will be utilized in analysis of the data. To integrate response accuracy with latency, the theory of the speed-accuracy tradeoff will be briefly considered.

Theory of Signal Detection

The Theory of Signal Detection (TSD) is essentially a statistical technique which permits the experimenter to determine the relationship between two states of the world--one in which signals are present and one in which they are not. This can then be compared with two possible responses of the subject. There are, therefore, four possible outcomes, a hit (the signal was present and the subject says it was), a correct rejection (there was no signal and the subject says there was not), a false alarm--F/A (there was no signal but the subject says there was), and a miss (there was a signal and the subject says there was not). A typical experiment using this technique would calculate the probability of hits and the probability of false alarms and map these into a ROC (Receiver Operating Characteristic) curve. (For details, see Egan, 1975). Both sensitivity (d′) and the subjective criterion (β) can be determined from the ROC.

The TSD by definition, therefore, looks at discrete events, the presence or absence of a signal within a given and well defined time interval. The paradigm, therefore, employs the situation in which uncertainty is at a minimum, i.e., the subject is externally paced by the experimenter and is asked for a response at the end of each time interval. The advantage of minimizing uncertainty is to enable the experimenter to compare time intervals in which the signal occurred with equal time intervals in which it did not.

The problem with this particular model is that while it is theoretically interesting its application is limited to a unique set of highly controlled laboratory experiments. Analysis of the types of problems that operators are confronted with in the real world, on the other hand, reveals a situation in which uncertainty is maximized. The

operator never knows when a signal could have occurred and is only paced by the existence of some internal trigger mechanism which responds or does not when some criterion has been met. Such an environment in particular characterizes the "task" of failure detection in aviation systems.

By definition, therefore, the application of TSD to this unique set of real-world circumstances poses a problem. The main problem lies in quantifying the false alarm rate so as to be able to compare it with the hit rate in some meaningful manner. The successful resolution of this problem would provide a meaningful measure of the effects of temporal uncertainty while still utilizing the unique advantages of TSD, in that separate measures of sensitivity and response bias would allow for a more fine-grained analysis.

The aim of this appendix is to examine the various approaches that have been developed to accomodate the maximization of the subjects' uncertainty in an unpaced responding task, and to show how, by adequate controls and changes, TSD can be profitably used in analyzing a real-monitoring task. By definition the studies on vigilance and the psychophysical studies using TSD have much in common in that they both measure the relative detectability of a signal. Where these research paradigms differ is in the observation period. By definition, the observation period is longer in the vigilance paradigm. For this reason, the early work did not use a TSD type analysis. However, ever since the application of TSD to this area, the vigilance literature has undergone a revision, and in some cases this has involved serious reinterpretation of some well established data. The main area of interest was the well documented deterioration in vigilance with the passing of time. "The majority of vigilance studies employing TSD

analysis have found that this decrement is prinicpally due to an increase in the strictness of the response criteria adopted by subjects, rather than to any loss of sensitivity over time" (Parasuraman & Davies, 1976) as had been previously thought.

Another problem with the classic vigilance work (Swets & Kristofferson, 1970) was its reliance on a dependent measure related solely to the frequence of detection, ignoring the well documented importance of latencies (Buck, 1966; Emerich et al., 1972; Parasuraman & Davies, 1976) in the detection paradigm. Parasuraman and Davies demonstrated how one can obtain different latencies for hits, false alarms, and correct rejections. The importance of this work lies in its demonstration that any accurate analysis of results must take into account and control for the latency of response in interpreting the results. This requirement, therefore, places special emphasis on the use of the Speed Accuracy Operating Characteristic (SAOC) discussed in the following section.

In discussing the application of TSD to vigilance-type tasks, Swets and Kristofferson (1970) point to the fact that both β and d' have been found to be sensitive to: (a) the low rates of false alarm (F/A) typically found in vigilance experiments, (b) violations or assumptions that underly the tabulated values of d' and β (this becomes particularly problematic when the signal rate is very low), (c) uncertainty about the nature of the signal. Taylor (1967) pointed out that the subjects uncertainty about what he is to respond to leads to asymmetric ROC curves. This variable is also sensitive to low signal rates. While Swets and Kristofferson (1970) point to a return to a more classic use of the TSD in vigilance studies, this solution is not always possible, especially in research designed to simulate real-world situations and

thereby maximize the uncertainty factor.

The development of two main methodologies has enabled the utilization of TSD in real-world settings: the method of free response (MFR) and the method of undefined intervals.

The Method of Free Response

As pointed out in the Introduction the real world cannot be easily divided into a series of equal time intervals which are then employed for comparing S's response to the occurrence or non-occurrence of a signal. We now have the unique situation in which the subject alone generates a series of responses, some hits and some false alarms (F/As). Luce (1966) saw, in this situation, a built-in ambiguity in that F/As do indeed occur and, furthermore, the latency of the responses to the hits had some variability. In order to determine causality, i.e., that the response was indeed made to the stimulus and not just another F/A, one has to make an arbitrary decision about the latency of the S-R interval. It was to this problem that Egan, Greenberg, and Shulman (1961a, 1961b, 1961c) addressed themselves in developing an analysis technique that they labelled the method of free response (MFR).

Egan et al. conducted a series of experiments and manipulated the subject's subjective information about the onset of the signal. In the first experiment (Egan et al., 1961a) the subject was not told when the signal would occur, only that it could occur anywhere within a defined interval which was always much longer than the signal itself. In the second experiment (Egan et al., 1961b) the subject was informed by the experimenter when the stimulus would occur. The third experiment (Egan et al., 1961c) was a more typical vigilance type experiment in that the signal was presented at any random time within the total observation interval. The time interval for these trials was two minutes, a short

period in terms of the vigilance type paradigm but much larger than the period used in the standard TSD model.

The successful evaluation of the temporal uncertainty situation characterizing the vigilance paradigm involves the ability to distinguish "yes" responses that are hits from "yes" responses that are false alarms. The method employed by Egan et al. (1961c) utilized frequency counts of responses immediately following the occurrence of the signal (D) as opposed to responses that occurred at longer latencies after the onset of the signal (O). The rates of response are cummulated and provide the estimates of the probability of hits as opposed to false alarms. Once such probabilities are known, ROC curves can be calculated.

It can be seen, therefore, that Egan et al. utilize the only data available--the latency data--to make inferences about the relative occurrence of hits as opposed to F/As. Their model is based on a number of fundamental though questionable assumptions.

(a) The subject divides time into a succession of temporal intervals T , and he mades a "yes-no" decision after each of these subjective intervals.

(b) The value of T is invariant with a change in the criterion adopted by the listener.

(c) There is a small variability of reaction time in the response so that the average reaction time for a given subject is irrelevant.

(d) The operating characteristics of the subject is best described by a power function.

As Green and Swets (1966) and Watson and Nichols (1976) point out, the success of the above model lies in its ability to predict the data, though much of the support for the above set of assumptions is by implication rather than direct proof. For example, the model assumes that as (the subject's subjective criterion) changes, the number of responses will change, and also that will be different for hits as opposed to F/As. Indeed it is the power law that defines the relationship between these two variables. This assumption gains support from the data presented by Egan et al. (1961c).

There are, however, a number of weaknesses with this model. As watson and Nichols (1976) point out "... the most serious shortcoming is that the measures of D and O are not estimated by identical procedures ... estimating D immediately after the signal, and then estimating O at a fixed time after the signal, always allows a potential bias in the estimate of O." (Watson & Nichols, 1976).

Another fundamental weakness of the model of Egan et al. lies in their assumption that a basic temporal interval (T) exists and that this interval remains constant throughout the trial. This assumption has great parsimonious value but has never been experimentally documented; indeed work on decision making and motor responding (Keele, 1976) leads to the conclusion that this interval is everchanging. The importance of the Egan et al. studies and the method of free response lies in the demonstration of the applicability of TSD type analysis to the free response paradigm, in which uncertainty is maximized.

<u>Method of Undefined Intervals</u>

In choosing a method that would come to grips with many of the problems raised above, Watson and Nichols (1976) asked the basic question, "Can an experiment be conducted in such a way that it is a

method of free response from the listener's point of view but still has the features of a defined-trial procedure essential to the reduction of the data into separate measures of sensitivity and response bias?" Their method involves the division of the total listening period into a number of idscrete and equal "measurement intervals." The signals to be detected are then randomly presented during 50% of the intervals while only background noise is presented on the remaining. These authors claim that this design insures that "the response criteria (will) remain the same, on the average, when "hits" and "false alarms" occur."

A further refinement employed by Watson and Nichols was the use of response latencies following the onset of a measurement interval, and in this way they obtained response latencies for the detection signals and for false responses to noise, F/A. This particular methodology combines the advantages of the MFR technique while obtaining estamates of the subject's sensitivity and bias that are independent of response rates. Naturally, their procedure may be modified by reducing the proportion of signal intervals to reproduce more accurately a vigilance situation.

Latency-Accuracy Analysis

The preceding discussion describes a procedure whereby, upon making certain assumptions, a reasonably bias-free estimate of detection accuracy may be obtained in paradigms, such as those employed in failure detection studies, when the observation interval is ill defined. It may be asserted on the basis of experimental literature, however, that detection accuracy represents only one dimension of performance in a monitoring situation (Buck, 1966). A well-known characteristic of decision making tasks is the operator's ability to trade off the speed of decision making for its accuracy. This speed-accuracy tradeoff has been studied extensively in reaction time paradigms (Pachella, 1974;

Fitts, 1966; Pew, 1969) and also to a lesser extent in paradigms of signal detection (Pike, 1973; Parasuraman & Davies, 1976). In any effort to compare detection "performance" across conditions then, the joint implications of speed and accuracy must be taken into account. For example, a condition that produces a high accuracy of responding might do so at such a prolonged latency that the utility of that decision in a real-world contest is less than that of a more rapid decision with slightly lower expected accuracy.

The experimental results described in the Results section are presented in the form of a joint speed-accuracy measure plotted in a space such as that depicted in Figure A-1. "Good" performance is represented by points lying in the upper left, in the region of fast accurate responses. Performance may be quantified by projecting the point locus obtained in any given condition onto a performance axis that runs from lower right to upper left. Experimental manipulations that shift performance parallel to this axis clearly affect the sensitivity to failures. Manipulations that shift performance orthogonally to the axis on the other hand represent shifts in a bias factor dictating fast inaccurate vs slow accurate responding. The units assigned to the performance index are clearly arbitrary but do require that an assumption be made with regard to the relative weighting of accuracy vs latency in detection. This weighting defines the scaling along the two axes or, equivalently, the slope of the performance axis. The exact numerical weighting used in the experiment is discussed on page 30.

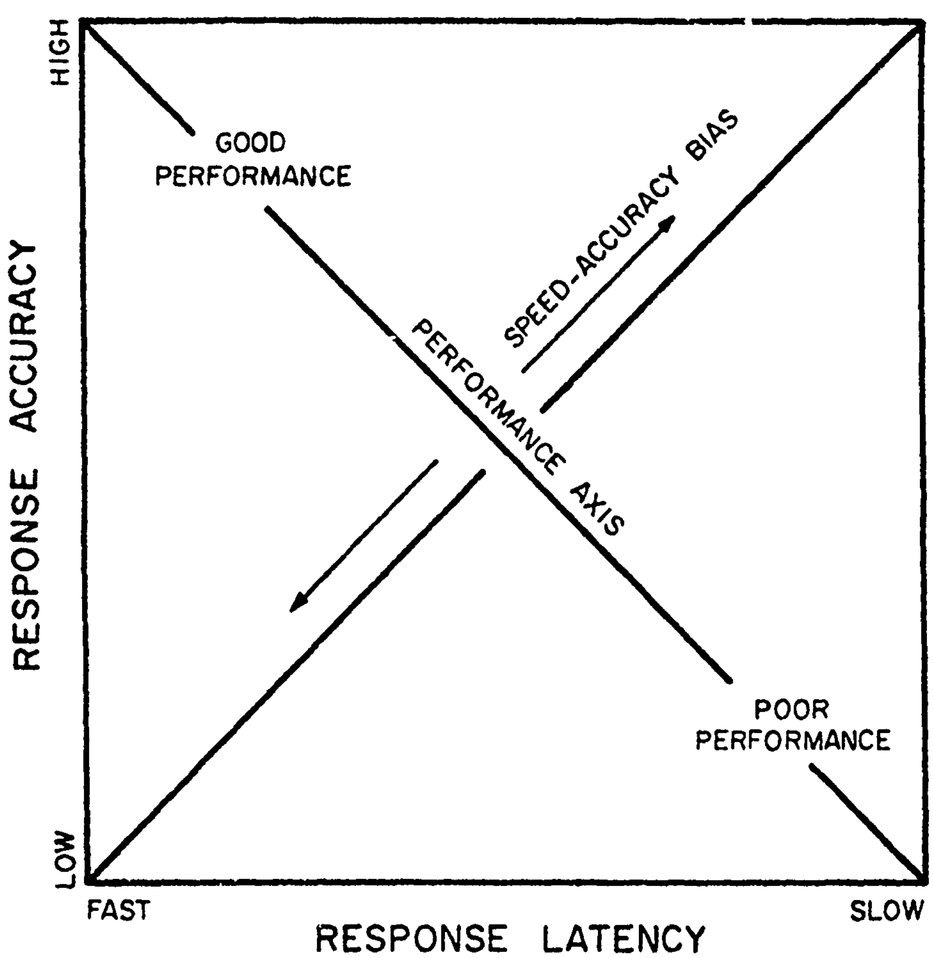


Figure A-1. Speed-Accuracy representation of detection performance.